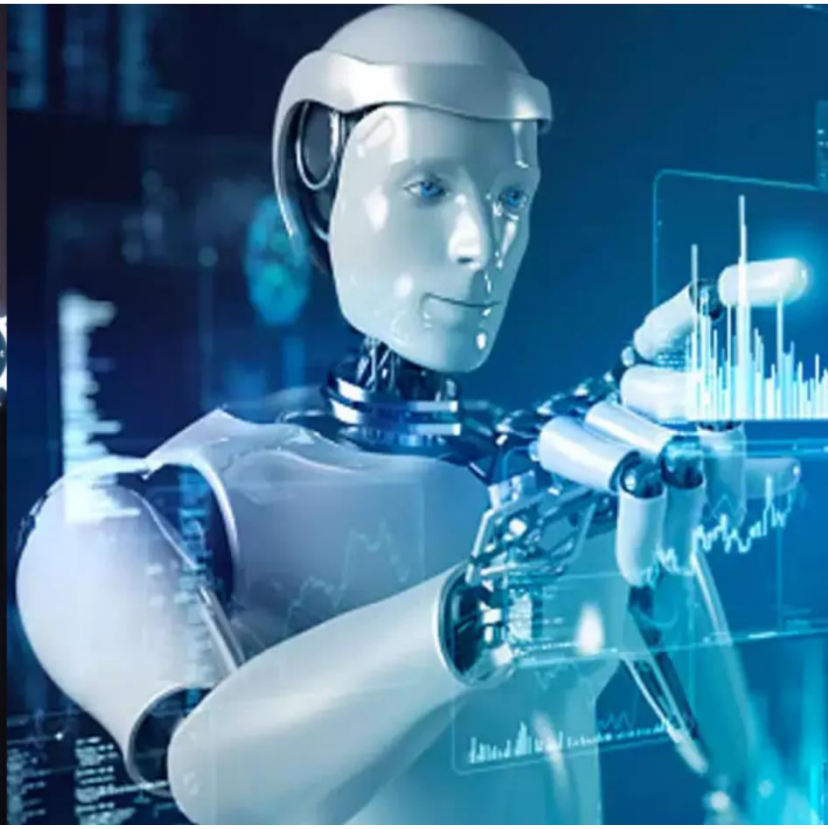
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

Churnguard: An Intelligent Model to Detect Potential Customer Loss

Abhishek Laharia, Sayed Azhar, Suraj Jha

Asst.Professor, Department of Electronics and Computer Science, Shree L.R. Tiwari College of Engineering, Mumbai, India

UG-Students, Department of Electronics and Computer Science, Shree L.R. Tiwari College of Engineering, Mumbai, India

ABSTRACT: Customer churn, the phenomenon of customers discontinuing their service or subscription, poses a significant challenge to businesses across various industries. Retaining existing customers is often more cost-effective than acquiring new ones, making accurate churn prediction a crucial task for proactive customer relationship management. This abstract explores the importance of customer churn prediction and highlights the application of various data mining and machine learning techniques to identify customers at high risk of leaving. By leveraging historical customer data, including demographics, usage patterns, and engagement metrics, and interaction history, predictive models can be built to forecast churn probability. These models enable businesses to implement targeted intervention strategies, such as personalized offers, proactive support, and improved service, to mitigate churn and enhance customer loyalty. This abstract will further discuss the challenges associated with churn prediction, including imbalanced datasets and the dynamic nature of customer behaviour, and briefly touch upon the evaluation metrics used to assess the performance of churn prediction models. Ultimately, effective customer churn prediction empowers organizations to make data-driven decisions, optimize customer retention efforts, and improve overall business profitability.

KEYWORDS: Customer Churn Prediction, Machine Learning Classification, Feature Engineering, Model Evaluation Metrics, Customer Retention

I. INTRODUCTION

The escalating costs associated with customer acquisition have underscored the critical importance of retaining existing customers for long-term business success. Customer churn, defined as the voluntary or involuntary termination of a customer's relationship with a company, directly impacts revenue streams and erodes profitability. While understanding past churn events provides valuable insights, the ability to predict which customers are likely to churn in the future offers a significant competitive advantage. This is where the power of machine learning comes into play, providing sophisticated analytical capabilities to transform raw customer data into actionable predictive intelligence. By employing various machine learning algorithms, businesses can move beyond reactive churn management to proactive intervention, fostering stronger customer loyalty and optimizing resource allocation.

Machine learning algorithms excel at identifying complex, non-linear relationships within large datasets that traditional statistical methods might overlook. In the context of customer churn, these algorithms can analyse a multitude of features, including demographic information (age, location, contract type), service usage patterns (data consumption, call frequency, feature utilization), engagement metrics (website visits, app usage, interaction history), customer sentiment (gleaned from reviews or support interactions), and billing details. By learning from historical instances of churn and non-churn, these algorithms can discern subtle patterns and indicators that collectively contribute to a customer's propensity to leave. This ability to process and interpret diverse data points with high accuracy makes machine learning an indispensable tool for building robust churn prediction models. The application of machine learning in churn prediction typically involves several key stages. First, a comprehensive dataset of customer information is collected and pre-processed, addressing issues such as missing values, outliers, and data inconsistencies. Feature engineering plays a crucial role in creating relevant and informative features that can enhance the predictive power of the models. Subsequently, various machine learning algorithms, such as logistic regression, decision trees,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

random forests, support vector machines, and gradient boosting techniques, are trained on historical data to learn the patterns associated with churn. The performance of these models is rigorously evaluated using appropriate metrics like accuracy, precision, recall, F1-score, and AUC (Area under the ROC Curve) to ensure their reliability and effectiveness in identifying potential churners. The ultimate goal is to deploy a model that can accurately flag at-risk customers, enabling businesses to implement timely and targeted retention strategies. These strategies might include offering personalized discounts, providing enhanced customer support, addressing specific pain points, or proactively engaging with dissatisfied customers to rebuild their loyalty and prevent them from switching to competitors. Furthermore, the dynamic nature of customer behaviour necessitates continuous monitoring and refinement of churn prediction models. Customer preferences, market trends, and competitive landscapes are constantly evolving, which can impact the factors driving churn. Therefore, it is crucial for businesses to regularly retrain their models with the latest data to maintain their accuracy and effectiveness over time. Additionally, understanding the "why" behind churn is as important as predicting "who" will churn. Techniques like feature importance analysis from machine learning models can provide valuable insights into the key drivers of customer attrition, allowing businesses to address underlying issues in their products, services, or customer experience to improve overall retention rates. In conclusion, machine learning offers a powerful and adaptable framework for customer churn prediction, empowering organizations to proactively manage customer relationships, reduce revenue loss, and foster long-term customer loyalty in an increasingly competitive marketplace.

II. LITERATURE SURVEY

The field of customer churn prediction has witnessed significant advancements with the increasing application of machine learning techniques. A review of the existing literature reveals a dynamic landscape of methodologies, data sources, and evaluation metrics aimed at accurately identifying customers at risk of leaving. Early research often focused on traditional statistical methods such as logistic regression. However, the growing availability of large and complex customer datasets has paved the way for more sophisticated machine learning algorithms. Decision trees and their ensemble variations, like Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM), have demonstrated robust predictive power due to their ability to capture non-linear relationships and handle feature importance effectively. Support Vector Machines (SVMs) have also been explored for their capability in highdimensional spaces, particularly when dealing with diverse customer attributes.

More recently, deep learning architectures, including Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs), are gaining traction, especially in scenarios where temporal patterns in customer behaviour (e.g., website visits, purchase history over time) are crucial indicators of churn. These models can automatically learn complex feature representations from raw data, potentially uncovering subtle churn signals that might be missed by traditional feature engineering approaches. An interesting trend in the literature is the emphasis on addressing the class imbalance problem, which is inherent in churn prediction datasets (where the number of non-churners significantly outweighs the number of churners). Various techniques, such as oversampling minority class (e.g., SMOTE), under sampling the majority class, and using cost sensitive learning approaches, have been investigated to mitigate the bias towards the majority class and improve the model's ability to accurately identify churners.

Furthermore, the literature highlights the importance of feature engineering and selection in building effective churn prediction models. Researchers have explored a wide range of features, including not only basic demographic and usage data but also more advanced features derived from customer interactions (e.g., sentiment analysis of customer reviews, call centre interactions), network characteristics (in telecommunications), and even macroeconomic indicators. The selection of the most relevant features through techniques like dimensionality reduction and feature importance analysis is crucial for model performance and interpretability.

Evaluation of churn prediction models goes beyond simple accuracy. Metrics such as precision, recall, F1-score, and Area under the Receiver Operating Characteristic curve (AUC-ROC) are commonly used to provide a more comprehensive assessment of the model's ability to correctly identify churners and avoid misclassifying non-churners. The choice of evaluation metric often depends on the specific business context and the relative costs of false positives and false negatives. Emerging research directions include the development of more interpretable machine learning models for churn prediction, allowing businesses to understand the reasons behind predicted churn and implement more targeted interventions. Additionally, there is growing interest in leveraging real-time data and streaming analytics to

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

predict churn in a more dynamic and timely manner. The integration of machine learning with explainable AI (XAI) techniques is also gaining prominence, aiming to build trust and facilitate better decision-making based on churn predictions.

III. PROBLEM STATEMENT AND OBJECTIVE

1. Develop a predictive model to identify customers at high risk of churn, enabling proactive intervention strategies to improve customer retention. This problem statement focuses on the practical application of the model in a business context, aiming to reduce customer attrition through targeted actions.
2. Investigate and compare the performance of various machine learning algorithms in accurately predicting customer churn based on available customer data, and determine the key factors that contribute most significantly to churn. This problem statement emphasizes the analytical and comparative aspects of the project, focusing on model performance and the identification of crucial churn drivers.

IV. METHODOLOGY

This work aims to predict customer churn in a commercial bank as early as possible using efficient data mining methods. A diagrammatic representation of the proposed model is given in Fig 1. Customer churn prediction using machine learning typically involves a structured, iterative process encompassing several key phases. This section outlines a comprehensive approach to building and deploying effective churn prediction models.

1. Data Acquisition and Understanding:

- Data Collection: The first step involves gathering relevant data from various sources within the organization. This may include:
- Customer Demographics: Age, gender, location, marital status, etc.
- Service/Product Usage: Subscription details, usage frequency, data consumption, feature utilization, purchase history, etc.
- Engagement Metrics: Website visits, app usage, interaction with marketing campaigns, social media activity, etc.
- Customer Interactions: Call centre logs, support tickets, chat transcripts, feedback surveys, etc.
- Billing Information: Payment history, plan changes, outstanding balances, etc.
- Business Understanding: Collaborating with domain experts to understand the business context of the data and identify potentially relevant features and business rules related to churn.

2. Data Pre-processing and Feature Engineering:

- Data Cleaning: Handling missing values (imputation or removal), identifying and treating outliers, and correcting any data inconsistencies or errors.
- Data Transformation: Converting categorical features into numerical representations (e.g., one-hot encoding, label encoding). Scaling numerical features (e.g., standardization, normalization) to ensure that algorithms are not biased towards features with larger ranges.
- Feature Engineering: Creating new features from existing ones that might provide more predictive power. This could involve:
- Aggregation: Calculating summary statistics (e.g., average usage, total spending) over specific time periods.
- Interaction Features: Combining two or more features to capture potential synergistic effects.
- Lag Features: Creating time-shifted versions of features to capture trends and temporal dependencies.
- Ratio Features: Deriving meaningful ratios from existing features (e.g., number of support tickets per usage).
- Feature Selection/Dimensionality Reduction: Identifying the most relevant features for the model and reducing the dimensionality of the dataset to improve model performance, reduce complexity, and mitigate the curse of dimensionality. Techniques include statistical tests, feature importance from tree-based models, and dimensionality reduction algorithms like Principal Component Analysis (PCA).

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

3. Model Development and Selection:

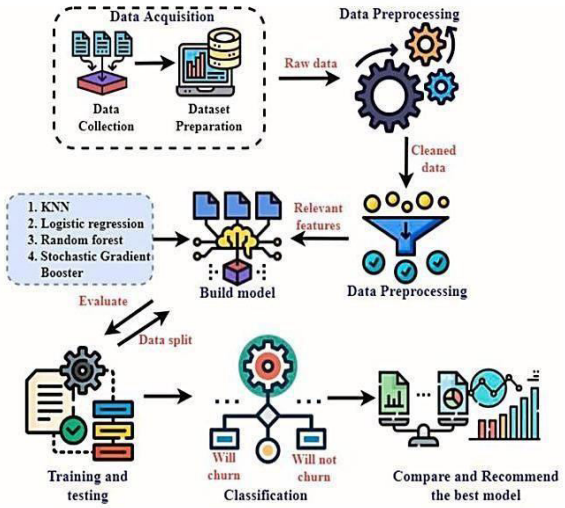


Fig. 1. System layout.

- Data Splitting: Dividing the pre-processed data into training, validation, and testing sets. The training set is used to train the model, the validation set is used to tune hyper parameters and compare different models, and the testing set is used for the final evaluation of the chosen model's performance on unseen data.
- Logistic Regression: A linear model suitable for binary classification.
- Decision Trees: Tree-based models that are easy to interpret.
- Random Forests: An ensemble of decision trees that often provides higher accuracy and robustness.
- Gradient Boosting Machines (e.g., XGBoost, LightGBM): Powerful ensemble methods known for achieving state-of-the-art results. Support Vector Machines (SVMs): Effective in high dimensional spaces.

4. Model Evaluation and Interpretation:

- Performance Metric Selection: Choosing appropriate evaluation metrics based on the business objectives and the characteristics of the churn problem. Common metrics include:
- Accuracy: Overall correctness of the model.
- Precision: Proportion of correctly predicted churners out of all predicted churners.
- Recall (Sensitivity): Proportion of actual churners correctly identified by the model.
- AUC-ROC: Area under the Receiver Operating Characteristic curve, which measures the model's ability to distinguish between churners and non-churners. .

5. Model Deployment and Monitoring:

- Deployment: Integrating the chosen model into the business's operational systems. This could involve: Batch Scoring: Periodically scoring the entire customer base to identify at-risk customers.
- Real-time Scoring: Scoring new customer data or changes in existing customer behaviour in real-time.
- API Integration: Exposing the model as an API for other applications to consume predictions.

6. Iteration and Refinement:

The churn prediction process is often iterative. Based on the model's performance, business feedback, and changes in data patterns, the methodology may involve revisiting earlier stages, such as feature engineering, model selection, or hyper parameter tuning, to further improve the accuracy and effectiveness of the churn prediction system. This structured methodology provides a robust framework for developing and deploying machine learning models for customer churn prediction, enabling businesses to proactively manage customer relationships and reduce attrition

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

TABLE III PREPROCESSED DATASET

Feature Name	Min	Max	Mean	SD
Credit Score	350	850	650.5288	96.6533
Gender	0	1	0.5457	0.4979
Age	18	92	38.9218	10.4878
Tenure	0	10	5.0128	2.8922
Balance	0	250898.09	76485.8893	62397.4052
Num Of Products	1	4	1.5302	0.5817
Has Cr Card	0	1	0.7055	0.4558
Is Active Member	0	1	0.5151	0.4998
Estimated Salary	11.58	199992.48	100090.2399	57510.4928

7. over Sampling

In data processing, oversampling and under sampling are strategies used to configure the class distribution of given data. Since the data is highly imbalanced (7963 positive class samples and 2037 negative class) and the size of the available data sample is small, this study will make use of the oversampling technique. Because if under sampling is preferred, the size of data will decrease in a way that enough data will not be there to build the model. Hence, this study is using the random oversampling by resampling the minority class (negative class).

8. Classification

The classification methods were applied over the pre-processed data. KNN, SVM, Decision Tree (DT) and RF classifiers are used for comparison of results. And also the comparison of results of different classifiers has been carried out over the selected features by different feature selection methods.

- K-Nearest Neighbour (KNN): The KNN method is one of the easiest and most efficient non-parametric ways of classification, based on supervised learning [12]. KNN works by identifying the k nearest samples from an existing dataset and when a new unknown sample appears, classify the new sample in the most similar class. That is, the classification algorithm determines the test sample group by the k training samples that are the nearest neighbours to the test sample and assign it to the class with the highest likelihood.
- Support Vector Machine (SVM): Support Vector Machine is an efficient, supervised machine learning algorithm derived from Vapnik's theory of statistical learning [13], [14], [15]. This has proved its success in the fields of classification [16], regression [17], time series prediction, and estimation in geotechnical practice and mining science [18]. SVM's main objective is to find an efficient distinguishing hyper plane that precisely categorizes data points and as far as possible, and distinguishes the points of two classes by reducing the possibility of misclassifying the training samples and unknown test samples. This implies that there is the maximum distance between two classes and the separating hyper plane. The Linear support vector machine (LSVM) model is used in this work. LSVM was originally developed to deal with binary class problems [19].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

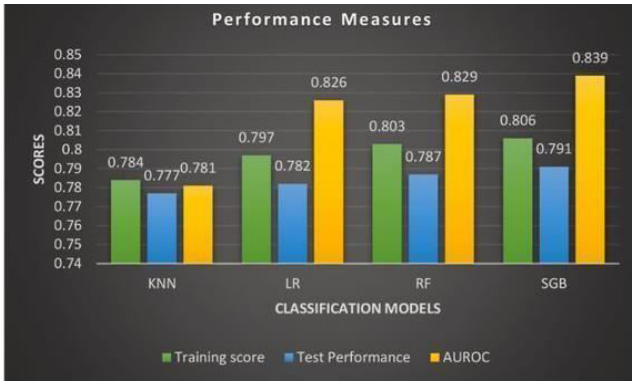
Impact Factor 8.699

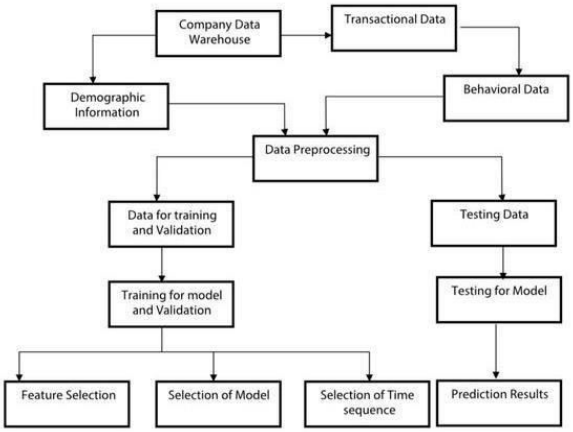
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

RF	81.75	92.19





9. Decision Tree (DT)

A decision tree is a procedure that slices a collection of data into various branch-like segments [20]. A tree of decisions is easy to read. This advantage makes explanations for the model simple. While another algorithm (like a neural network) can generate a much more accurate model in a given scenario, a decision tree could be trained to predict the neural network's predictions, thus opening up the neural networks "black box". Another benefit is that, in the correlation between the target variables and the predictor variables it can model a high degree of nonlinearity. A decision tree is composed of two major strategies [21]; Tree creation and Classification.

10. Random Forest (RF)

Breiman [22] presented RF as an ensemble classifier for tree learners. The method employs several decision trees so that each tree relies on the values of an individually selected random vector with the same distribution for all trees. Right choice for the tendency of decision trees to overfit their training collection. In short, Random forests are actually a way to combine many deep decision trees which are learned on various sections of the same dataset with the target of decreasing the variance. The real advantage of using RF is it comes with quite high dimensional data, with no need to perform dimensionality reduction and feature selection. The training rate is also higher and ease to use in parallel models.

V. EXPERIMENTAL RESULTS

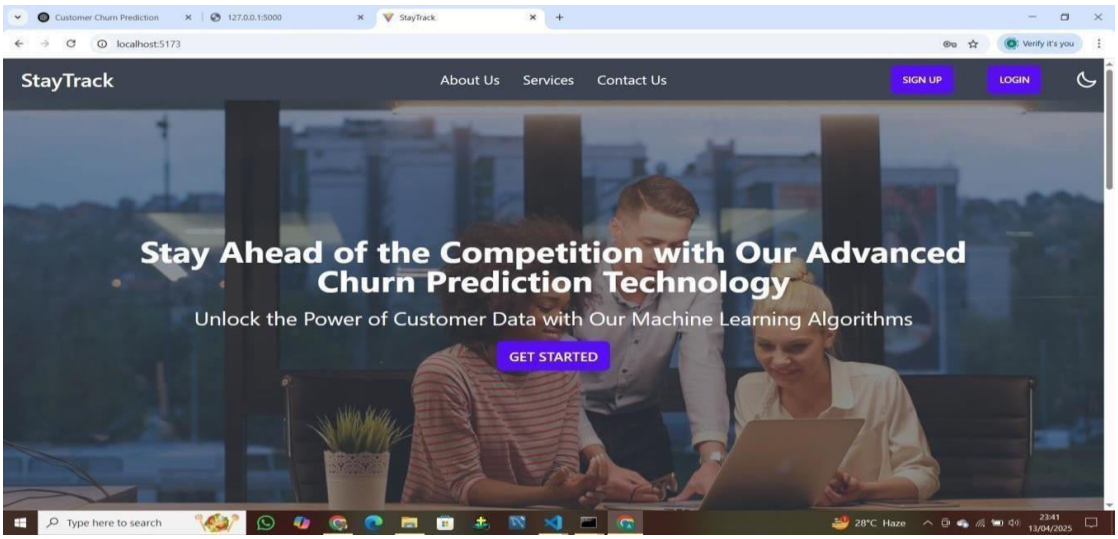


Figure 4.1 Stay Track (Web App) Interface

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

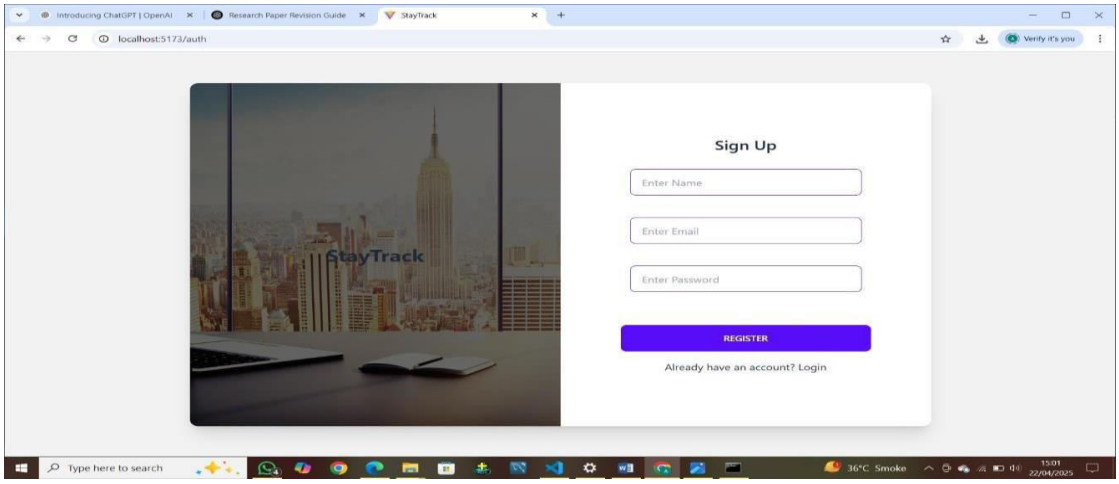


Figure 4.2 Sign up Page

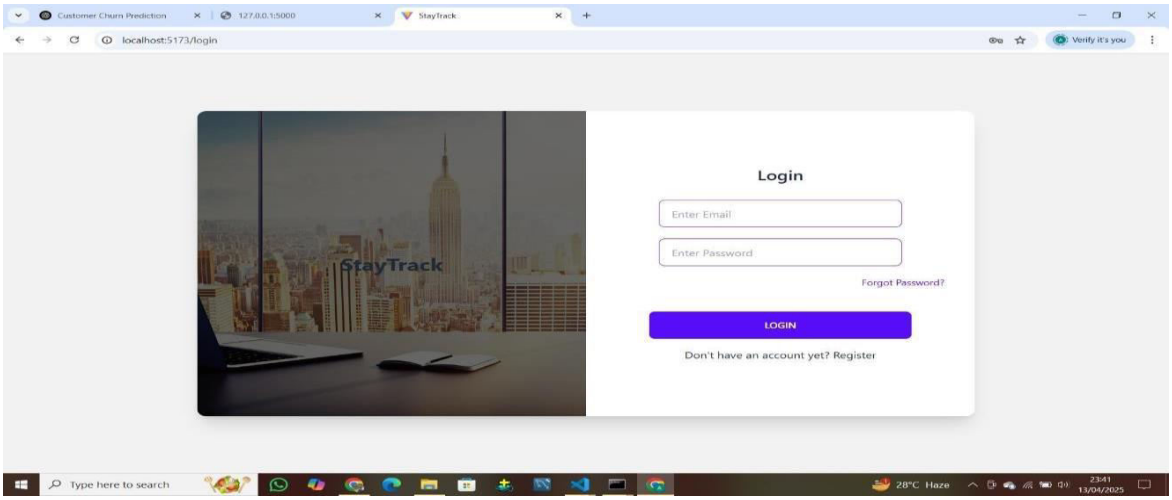


Figure 4.3 Login Page

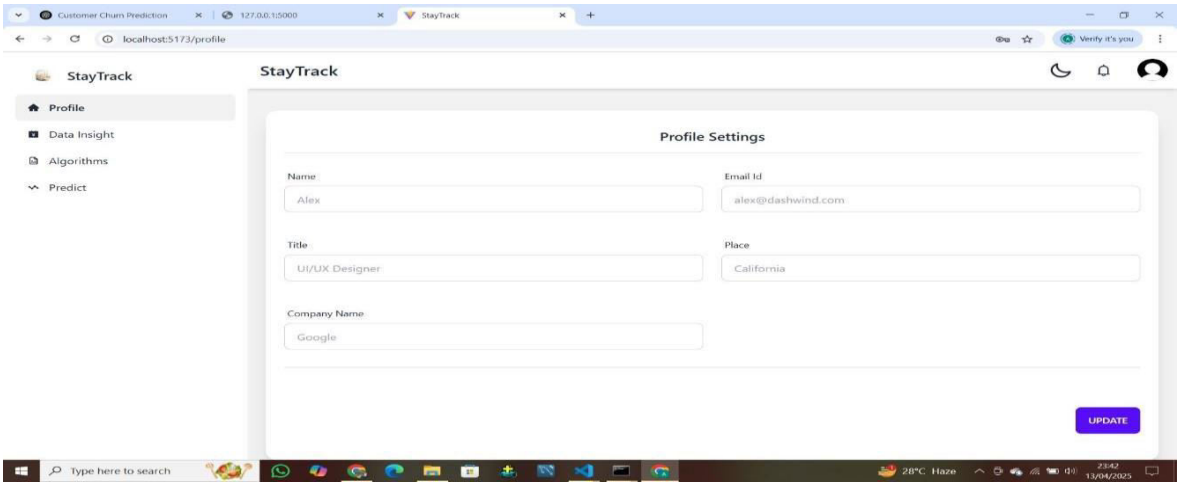


Figure 4.4 Profile Page

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

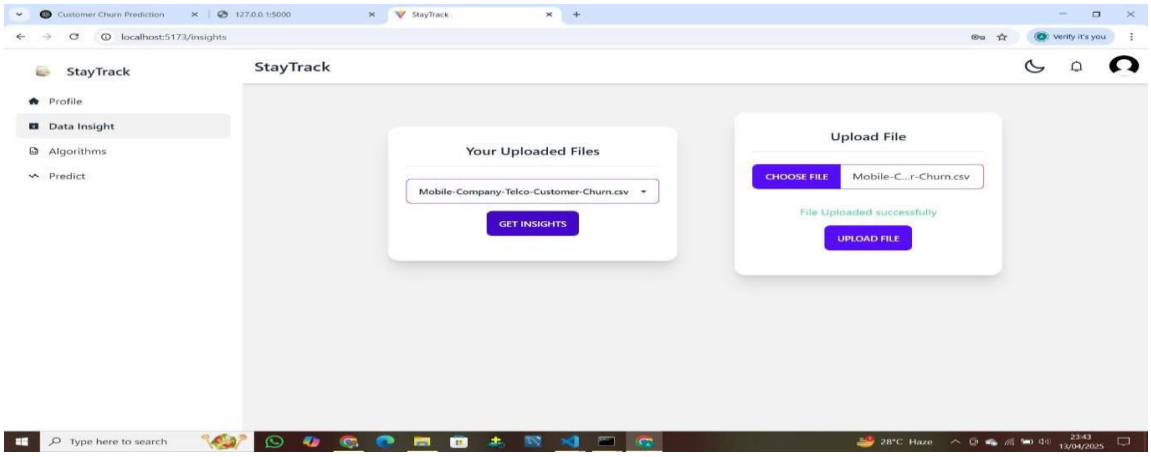


Figure 4.5 Data Insights Tab

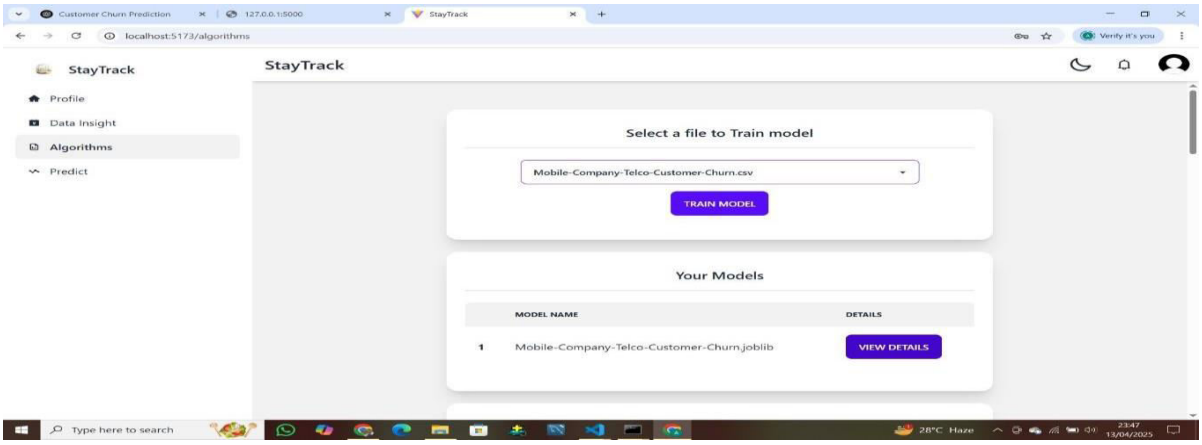


Figure 4.6 Algorithms Tab

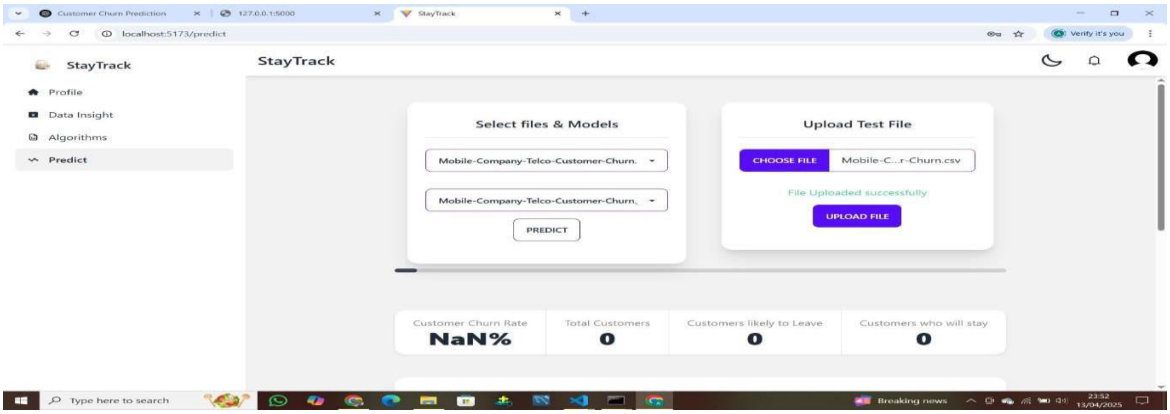


Figure 4.7 Predict Tab

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404648|

VI. CONCLUSIONS

While the banking sector is considered, like any other organization, customer engagement has become one of the primary concerns. To resolve this crisis, banks need to identify customer churn possibilities as quickly as possible. There are various studies ongoing in banking churn prediction. Different entities measure the churn rate of customers in various ways using different bits of data or information. The need for a system that can forecast the client churning in banking in a generalized way in the early stages is really important. The system needs to works with fixed and potential data sources that are independent of any service provider. And also the model must be in a form in which; can use minimal information and can give maximum throughput for the prediction. This study focus to fulfil these needs. The purpose of this study is to build the most appropriate model to predict client churn in a Bank in the early stages. The study only used a small amount of data (10000 samples), and also highly imbalanced. But real commercial bank data would be much larger. By oversampling, both of these headaches up to a certain degree can be resolved. The model examined KNN, SVM, Decision Tree, RF classifiers under different conditions for this study. A better result is achieved when using the RF classifier together with oversampling (95.74%). Feature selection methods have nothing to do with tree classifiers (Decision Tree and RF). As the result indicates, feature reduction (feature selection) is decreasing the prediction score of tree classifiers. Another observation is that unlike other classifiers, in SVM, oversampling is decreasing the score. It's because the Bank dataset is lopsided. Hence, SVM unable to handle the data well enough.

REFERENCES

1. D.-R. Liu and Y.-Y. Shih, "Integrating ahp and data mining for product recommendation based on customer lifetime value," Information & Management, vol. 42, no. 3, pp. 387–400, 2005.
2. G. Canning Jr, "Do a value analysis of your customer base," Industrial Marketing Management, vol. 11, no. 2, pp. 89–93, 1982.
3. R. W. Stone and D. J. Good, "The assimilation of computer-aided marketing activities," Information & management, vol. 38, no. 7, pp. 437–447, 2001. [4] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," IEEE Transactions on Knowledge and data Engineering, vol. 8, no. 6, pp. 866–883, 1996.
4. M.-K. Kim, M.-C. Park, and D.-H. Jeong, "The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services," Telecommunications policy, vol. 28, no. 2, pp. 145–159, 2004.
5. T. J. Swamidason, "Survey of data mining algorithm's for intelligent computing system," Journal of Trends in Computer Science and Smart Technology, vol. 01, pp. 14–23, 09 2019.
6. B. He, Y. Shi, Q. Wan, and X. Zhao, "Prediction of customer attrition of commercial banks based on svm model," Procardia Computer Science, vol. 31, pp. 423–430, 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details