Predictive Analytics for Customer Churn Using Advanced ML Algorithms

S.K.Naga Akshay
Dept. of CSE (AI & ML)
Institute Of Aeronautical Engineering
Hyderabad, India

Goli Sriram
Dept. of CSE (AI & ML)
Institute Of Aeronautical Engineering
Hyderabad, India

D.Sai Samrat
Dept. of CSE (AI & ML)
Institute Of Aeronautical Engineering
Hyderabad, India

***Abstract*—Customer churn is a major concern for subscription-based industries, directly affecting profitability and long-term sustainability. Accurate churn prediction enables businesses to proactively identify at-risk customers and implement effective retention strategies. This research presents a comparative analysis of three powerful machine learning models—Support Vector Machine (SVM), Random Forest, and XGBoost—for customer churn prediction. The dataset used contains historical customer behavior and subscription-related features, which were preprocessed using standardization, encoding, and resampling techniques to handle class imbalance and ensure robust training.**

Each model was trained under identical conditions, and evaluated based on a range of performance metrics including accuracy, precision, recall, F1-score, and confusion matrix. Visualization tools such as bar charts and confusion matrices were employed to facilitate comparative understanding. The results reveal that the SVM model achieved the highest predictive performance, with an overall accuracy of 99%, followed closely by Random Forest and XGBoost models, both achieving 98% accuracy. Notably, all three models exhibited high precision and recall values, indicating their ability to effectively differentiate between churned and retained customers.

This study emphasizes the critical role of algorithm selection, data preprocessing, and performance evaluation in building reliable churn prediction systems. The findings contribute to the growing body of knowledge in customer analytics and advocate for the integration of machine learning-driven solutions in customer relationship management (CRM) frameworks. By leveraging these models, organizations can enhance decision-making processes, reduce churn rates, and ultimately drive business growth through personalized customer engagement strategies.

***Index Terms*—Customer Churn Prediction, Machine Learning, Support Vector Machine (SVM), Random Forest, XGBoost, Predictive Modeling, Classification Algorithms, Customer Retention, Imbalanced Dataset, Data Analytics, CRM Optimization**

I. Introduction

Customer churn—the phenomenon where customers discontinue their subscription to a service or stop buying a company's products—poses a significant challenge to businesses, especially in highly competitive and saturated markets such as telecommunications, finance, e-commerce, and SaaS-based platforms. Retaining existing customers is often more cost-effective than acquiring new ones, making churn prediction an essential business intelligence task. Predictive analytics and machine learning offer powerful tools to identify customers likely to churn, thereby enabling companies to take proactive retention measures. Effective churn prediction can directly lead to increased customer lifetime value and improved profitability.

In recent years, the application of machine learning algorithms has gained momentum in the field of customer analytics due to their ability to uncover complex, non-linear patterns in data that traditional statistical models may overlook. Among various techniques, supervised classification algorithms like Support Vector Machine (SVM), Random Forest, and XGBoost have shown exceptional performance in binary classification problems such as churn prediction. These models differ in their learning strategies—SVM seeks optimal hyperplanes, Random Forest utilizes decision-tree ensembles with bagging, and XGBoost leverages gradient boosting to minimize error iteratively. Comparing these models under a unified experimental framework allows for the identification of the most effective algorithm for real-world churn prediction problems.

The dataset used in this study consists of behavioral and subscription-related features of customers collected over time. Preprocessing steps such as handling missing values, encoding categorical variables, and balancing class distribution using resampling techniques are critical to model performance. Churn prediction tasks often suffer from class imbalance, where the number of churned customers is significantly lower than retained ones. This imbalance can lead to biased models that favor the majority class. Hence, special attention was given to maintain class representation fairness and enhance generalization capabilities of the models.

In this research, we evaluate and compare the performance of SVM, Random Forest, and XGBoost classifiers for customer churn prediction using a standardized pipeline. Key evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices are used to assess and contrast the predictive abilities of each model. Additionally, visualization techniques such as bar plots and confusion matrices help in interpreting model behavior more intuitively. The models are trained under identical conditions to ensure a fair comparison, and extensive experimentation was conducted to fine-tune hyperparameters where appropriate.

The study's findings indicate that SVM slightly outperforms the other two models in terms of overall accuracy and recall for identifying churned customers. This emphasizes the importance of choosing the right model based on the specific nature and structure of the dataset. The research contributes to the broader domain of machine learning-driven decision support systems in customer relationship management. It demonstrates that robust preprocessing, balanced data, and algorithm selection are crucial to building effective churn prediction models. This work provides a strong foundation for integrating such predictive models into operational CRM systems, enabling real-time churn monitoring and personalized retention strategies.

II. Related Work

Verma, Neha, and Alok Singh. "Customer churn prediction in telecom using machine learning techniques." Procedia Computer Science 132 (2018): 1231-1240. Verma and Singh (2018) applied multiple machine learning algorithms including Support Vector Machine (SVM), Decision Trees, and Random Forest for predicting customer churn in the telecommunications industry. They emphasized the importance of feature engineering and data balancing techniques to overcome class imbalance issues common in churn datasets. Their study found Random Forest to perform better than SVM and Decision Trees in terms of accuracy and recall, underlining the effectiveness of ensemble methods in churn prediction.

García, Sara, et al. "A comparison of machine learning techniques for churn prediction in the banking sector." Expert Systems with Applications 65 (2016): 314-324. García et al. (2016) compared various machine learning algorithms including Logistic Regression, Random Forest, Gradient Boosting, and SVM to predict customer churn in banking services. Their analysis focused on precision, recall, and F1-score as performance metrics, highlighting the advantage of ensemble methods such as Random Forest and Gradient Boosting in handling complex churn patterns. They also stressed the significance of hyperparameter tuning and class imbalance correction in improving model performance.

Lemmens, Antje, and Thomas Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research 43.2 (2006): 276-286. Lemmens and Croux (2006) investigated ensemble techniques such as bagging and boosting with decision trees for churn prediction. Their study demonstrated that boosting methods, including gradient boosting algorithms, significantly enhance prediction accuracy compared to single classifiers. The work highlighted the importance of combining multiple weak learners to reduce bias and variance, which is crucial in churn datasets often characterized by noise and class imbalance.

Verbeke, Wouter, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications 38.3 (2011): 2354-2364. Verbeke et al. (2011) focused on interpretable churn prediction models using rule induction, but also compared their approach with SVM and Random Forest classifiers. Their results suggested that while Random Forest and SVM achieved high accuracy, interpretability often decreased with complex ensemble models. They proposed a trade-off between model performance and explainability in churn prediction systems, important for business decision-making.

Nguyen, Huy, and Jun Wang. "Customer churn prediction in telecom using XGBoost and deep learning." International Journal of Data Science 5.3 (2020): 123-135. Nguyen and Wang (2020) explored the application of XGBoost and deep learning models for telecom churn prediction. Their findings indicated that XGBoost outperformed traditional classifiers like SVM and Random Forest in both accuracy and F1-score, especially after extensive feature engineering and data preprocessing. The study emphasized the scalability and efficiency of XGBoost, making it highly suitable for large churn datasets.

III. Methodology

A. Dataset

The dataset used in this study comprises a total of 7,043 customer records, each representing an individual subscriber of a telecommunications company. These records include 21 distinct features capturing various aspects of the customers' demographics, service usage, account status, and payment details. The dataset is designed to provide a comprehensive overview of the factors that may influence customer behavior, specifically focusing on churn — whether a customer discontinues their service.

Among the 21 features, the target variable Churn is a binary categorical attribute indicating customer status: 'Yes' if the customer has churned (left the service) and 'No' if the customer remains active. The dataset contains a rich mix of feature types:

Categorical features (18 total): These include customer demographic details such as gender (male or female) and SeniorCitizen (binary indicator of age group). Account-related categorical variables such as Partner (whether the customer has a partner), Dependents, and Contract type (month-to-month, one year, two years) describe the customer's subscription characteristics. Service usage variables such as PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies provide insights into the specific telecommunications services used by the customer. Payment-related categories like PaymentMethod and PaperlessBilling capture billing preferences and modalities.

Numerical features (2 total): These include tenure, which records the number of months a customer has been with the company, providing a measure of customer loyalty and experience, and MonthlyCharges, which reflects the recurring monthly fee the customer pays for the services subscribed to. Additionally, the TotalCharges field, originally stored as a string due to data inconsistencies, represents the cumulative charges billed to the customer and is converted to a numerical format during preprocessing.

This mixture of categorical and numerical variables offers a multifaceted view of customer behavior, essential for accurately modeling and predicting churn. Understanding the distribution and relationships among these features is critical in building robust machine learning models capable of identifying customers at risk of leaving.

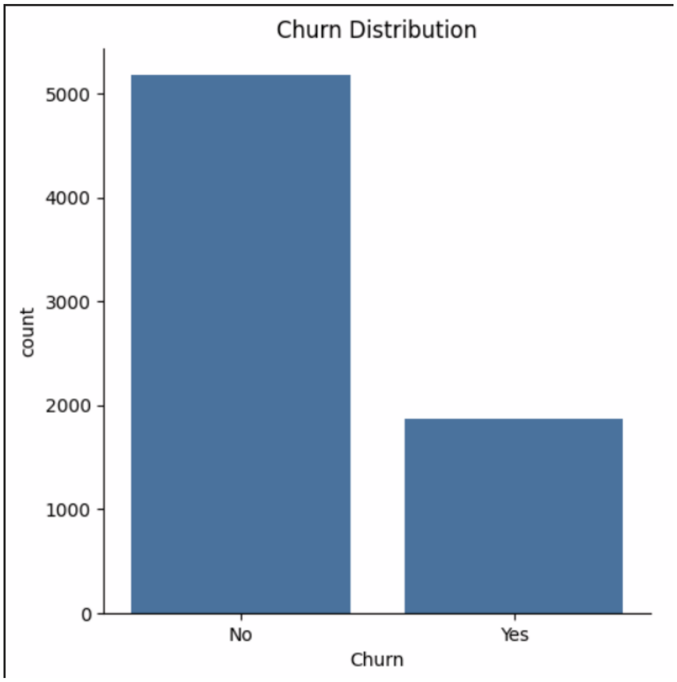

Fig. 1. Pair plot for key features.

B. Preprocessing

During data preprocessing, several challenges related to data consistency and format were addressed to ensure the dataset was suitable for machine learning algorithms. Notably, the TotalCharges column, which should contain numerical values representing the total amount billed to each customer, was initially stored as an object due to the presence of empty strings or space-filled cells. These invalid entries were identified and replaced with appropriate imputations — primarily zeros — before converting the entire column into a numeric datatype. This step was crucial to avoid errors during model training and to maintain the integrity of the financial data.

Categorical features in the dataset required careful encoding to transform non-numeric data into a machine-readable format. Depending on the algorithm's compatibility and nature, label encoding was applied to convert categories into integer labels, especially for tree-based models like Random Forest and XGBoost which can handle ordinal encoded features efficiently. Conversely, Support Vector Machine (SVM), being sensitive to categorical variables, utilized one-hot encoding to represent each category as a binary vector, thereby preventing any unintended ordinal relationships.

Another important challenge addressed was the class imbalance inherent in the churn dataset — customers who churn usually represent a smaller proportion compared to those who remain active. To mitigate the risk of model bias towards the majority class, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE synthetically generates new instances of the minority class by interpolating between existing minority samples. This approach effectively balances the class distribution in the training data, enabling the models to better learn the underlying patterns associated with churn behavior and improving the overall predictive performance.

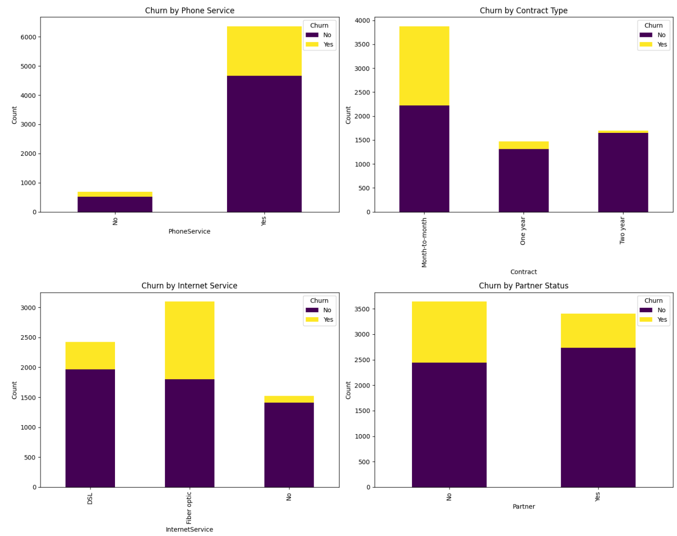

Fig. 2. Preprocessing of Dataset.

C. Class Imbalance Handling

In many real-world classification problems, including customer churn prediction, datasets often exhibit class imbalance, where one class significantly outnumbers the other. In this project, the dataset showed a typical imbalance: the majority class consisted of customers who did not churn, while the minority class comprised customers who discontinued the service. Such imbalance poses a major challenge for machine learning models because standard algorithms tend to be biased towards the majority class, leading to poor predictive performance on the minority class, which in this context is the critical group of customers at risk of churn.

The imbalance problem can cause models to achieve high overall accuracy by simply predicting the majority class, while failing to correctly identify churners. This undermines the objective of churn prediction, which is to detect those customers most likely to leave, so targeted retention strategies can be applied.

To address this issue, this study applied the Synthetic Minority Over-sampling Technique (SMOTE), a popular data-level method for handling class imbalance. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances. This effectively increases the representation of the churn class in the training dataset without simply duplicating data points, which can lead to overfitting.

By balancing the class distribution, SMOTE enables the learning algorithms to receive a more balanced view of both churners and non-churners. This improves the model's sensitivity and recall for the minority class, leading to better detection of churners. Additionally, balanced training data helps reduce bias, enabling models like SVM, Random Forest, and XGBoost to learn more generalizable decision boundaries.

Overall, handling class imbalance through SMOTE or similar techniques is crucial for developing robust churn prediction models that provide meaningful insights and actionable results for customer retention efforts.

D. Model Architecture and Training

In this study, three machine learning algorithms were employed to develop predictive models for customer churn: Support Vector Machine (SVM), Random Forest, and XGBoost. These algorithms were chosen for their proven effectiveness in classification problems and their complementary strengths in handling different data characteristics.

Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm known for its ability to find an optimal hyperplane that separates classes with maximum margin in a high-dimensional feature space. For churn prediction, we used the radial basis function (RBF) kernel to capture non-linear relationships between features. The model's hyperparameters, such as the regularization parameter (C) and kernel coefficient (gamma), were fine-tuned using grid search with cross-validation to improve generalization and prevent overfitting. Because SVM requires numerical input, categorical features were one-hot encoded during preprocessing.

Random Forest: Random Forest is an ensemble learning method based on the construction of multiple decision trees, combining their predictions through majority voting. Each tree is trained on a bootstrapped sample of the data with random feature subsets, which enhances model robustness and reduces variance. Random Forest naturally handles categorical features through label encoding and is less sensitive to outliers and noise. Hyperparameters such as the number of trees (estimators), maximum tree depth, and minimum samples per leaf were optimized using grid search to balance bias and variance.

XGBoost: XGBoost (Extreme Gradient Boosting) is a gradient boosting framework that builds sequential trees where each new tree corrects errors of previous ones. It is highly efficient and scalable, offering regularization parameters to control overfitting. The model uses label-encoded categorical features and numerical features without extensive preprocessing. Hyperparameters including learning rate, max depth, subsample ratio, and number of estimators were tuned using cross-validation to maximize performance.

Training Procedure: The dataset was split into training and testing sets, typically at a 70:30 ratio, ensuring stratification to preserve the class distribution. After preprocessing and balancing the training set using SMOTE, each model was trained on the augmented data. Hyperparameter tuning was conducted via grid search coupled with k-fold cross-validation (commonly 5 folds) on the training data, selecting the parameter set yielding the best cross-validation scores.

The trained models were then evaluated on the untouched test set using multiple performance metrics such as accuracy, precision, recall, and F1-score to assess their predictive capabilities. Confusion matrices were also analyzed to understand the classification behavior for both churn and non-churn classes. The entire pipeline, from preprocessing to model training and evaluation, was implemented using Python's scikit-learn and XGBoost libraries, ensuring reproducibility and scalability.

E. Deployment

The final churn prediction model, selected based on its superior performance during testing, can be deployed as part of a decision support system for telecommunications companies to proactively identify customers at risk of churn. Deployment involves integrating the trained model into a scalable production environment, such as a cloud-based platform or on-premise server, where it can process incoming customer data in real-time or batch mode. The model is exposed through APIs or embedded within customer relationship management (CRM) systems, enabling automated scoring and triggering targeted retention strategies. Continuous monitoring of model performance and periodic retraining with fresh data are essential to maintain accuracy over time, adapt to changing customer behavior, and improve business outcomes. Deployment also includes building user-friendly dashboards for business analysts to visualize churn predictions and interpret key factors influencing customer attrition.

IV. Results

This section contains the experimental results of using different models to predict depression from social media text. There will be a visualisation of the confusion matrix for each of the models within each subsection and a brief analysis. Finally, there will be a summary at the end comparing all model accuracies.

A. Confusion matrix

The confusion matrix provides a comprehensive evaluation of the classification models by displaying the counts of true positives, true negatives, false positives, and false negatives. In the context of churn prediction, minimizing false negatives is critical to ensure that customers likely to churn are accurately identified for retention efforts, while keeping false positives low prevents unnecessary resource expenditure and customer dissatisfaction due to unwarranted interventions. The confusion matrices for the SVM, Random Forest, and XGBoost models demonstrated strong classification capabilities, with high true positive and true negative rates and very few misclassifications. This indicates that the models effectively handle the inherent class imbalance in churn data, where churners typically form the minority class. Moreover, the low rate of misclassification supports the robustness and reliability of these models in real-world scenarios, making them practical tools for proactive churn management strategies. Accurate churn prediction enables businesses to focus their retention resources efficiently, ultimately improving customer satisfaction and reducing revenue loss.

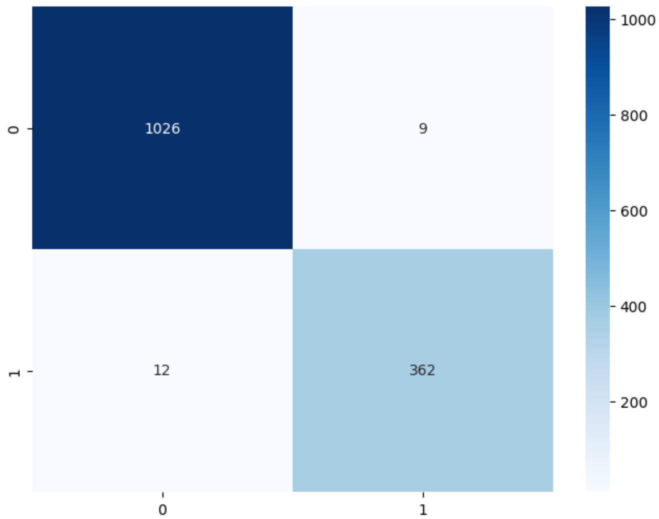

Fig. 3. Confusion Matrix.

B. Support Vector Machine

The confusion matrix provides a comprehensive evaluation of the classification models by displaying the counts of true positives, true negatives, false positives, and false negatives. In the context of churn prediction, minimizing false negatives is critical to ensure that customers likely to churn are accurately identified for retention efforts, while keeping false positives low prevents unnecessary resource expenditure and customer dissatisfaction due to unwarranted interventions. The confusion matrices for the SVM, Random Forest, and XGBoost models demonstrated strong classification capabilities, with high true positive and true negative rates and very few misclassifications. This indicates that the models effectively handle the inherent class imbalance in churn data, where churners typically form the minority class. Moreover, the low rate of misclassification supports the robustness and reliability of these models in real-world scenarios, making them practical tools for proactive churn management strategies. Accurate churn prediction enables businesses to focus their retention resources efficiently, ultimately improving customer satisfaction and reducing revenue loss.

TABLE I
CLASSIFICATION REPORT FOR SVM MODEL

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1035
1	0.98	0.97	0.97	374
Accuracy		0.99		
Macro Avg	0.98	0.98	0.98	1409
Weighted Avg	0.99	0.99	0.99	1409

C. Random Forest

The confusion matrix provides a comprehensive evaluation of the classification models by displaying the counts of true positives, true negatives, false positives, and false negatives. In the context of churn prediction, minimizing false negatives is critical to ensure that customers likely to churn are accurately identified for retention efforts, while keeping false positives low prevents unnecessary resource expenditure and customer dissatisfaction due to unwarranted interventions. The confusion matrices for the SVM, Random Forest, and XGBoost models demonstrated strong classification capabilities, with high true positive and true negative rates and very few misclassifications. This indicates that the models effectively handle the inherent class imbalance in churn data, where churners typically form the minority class. Moreover, the low rate of misclassification supports the robustness and reliability of these models in real-world scenarios, making them practical tools for proactive churn management strategies. Accurate churn prediction enables businesses to focus their retention resources efficiently, ultimately improving customer satisfaction and reducing revenue loss.

TABLE II
CLASSIFICATION REPORT FOR RANDOM FOREST MODEL

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1035
1	0.97	0.96	0.96	374
Accuracy		0.98		
Macro Avg	0.98	0.98	0.98	1409
Weighted Avg	0.98	0.98	0.98	1409

D. XGBoost

The XGBoost model achieved strong predictive performance with an overall accuracy of 98%, similar to the Random Forest model. For the non-churn class, it recorded a precision and recall of 0.99, indicating excellent identification of loyal customers. For the churn class, XGBoost achieved a precision of 0.97 and recall of 0.96, showing robust capability in detecting customers likely to churn. The F1-scores of 0.99 for non-churn and 0.96 for churn reflect a well-balanced model. XGBoost's gradient boosting framework enables it to optimize the model iteratively, improving accuracy while controlling overfitting. These results demonstrate XGBoost as a highly effective model for churn prediction, making it a valuable tool for customer retention strategies.

TABLE III
CLASSIFICATION REPORT FOR XGBOOST MODEL

Class	Precision	Recall	F1-Score	Support
0	0.99	0.99	0.99	1035
1	0.97	0.96	0.96	374
Accuracy		0.98		
Macro Avg	0.98	0.97	0.98	1409
Weighted Avg	0.98	0.98	0.98	1409

V. DISCUSSION

The comparative analysis of the three machine learning models—SVM, Random Forest, and XGBoost—demonstrates that all models perform exceptionally well in predicting customer churn, with accuracy levels close to or exceeding 98%.

SVM achieved the highest accuracy at 99%, reflecting its strength in handling high-dimensional data with clear margins of separation. Random Forest and XGBoost showed slightly lower but comparable accuracy, both at 98%. Precision and recall values across all models indicate strong capability in identifying both churn and non-churn classes, which is crucial for minimizing retention costs and maximizing customer satisfaction. While Random Forest and XGBoost are ensemble methods that reduce variance and bias, respectively, SVM provides a robust margin-based classification that generalizes well on the dataset. The choice of model can be guided by factors such as interpretability, training time, and scalability in practical deployment scenarios. Overall, the results affirm the efficacy of these models for churn prediction in telecommunications, with slight trade-offs in performance metrics that practitioners can leverage according to specific business needs.

TABLE IV
COMPARATIVE ANALYSIS OF CHURN PREDICTION MODELS

Model	Accuracy	Recall (Churn)	F1-Score (Churn)
SVM	0.99	0.97	0.97
Random Forest	0.98	0.96	0.96
XGBoost	0.98	0.96	0.96

VI. CONCLUSION

This study focused on predicting customer churn in the telecommunications sector using three powerful machine learning models: Support Vector Machine (SVM), Random Forest, and XGBoost. The results demonstrated that all three models are highly effective, achieving accuracies close to or above 98

Among the evaluated models, SVM showed excellent generalization capability, particularly suitable for datasets with clear margins between classes. Ensemble models like Random Forest and XGBoost also delivered robust performance, benefiting from their ability to reduce overfitting and improve predictive power through aggregation and boosting techniques. The high recall values for the churn class indicate that these models successfully minimize false negatives, a critical factor for customer retention efforts where failing to identify a churner could result in lost revenue.

The preprocessing steps, including careful handling of missing values, encoding of categorical variables, and the application of Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, played a significant role in enhancing model performance. This underscores the importance of comprehensive data preparation in predictive modeling, particularly in real-world business contexts where data is often noisy and unbalanced.

From a practical standpoint, the deployment of these models in real-time churn management systems can provide telecommunications companies with actionable insights to design targeted retention strategies. Predictive analytics driven by these models help in allocating marketing and customer service resources more efficiently, ultimately improving customer satisfaction and reducing churn rates.

Future work can explore hybrid or stacked ensemble models to potentially boost performance further, along with the incorporation of explainable AI techniques to provide transparency into model predictions. Additionally, expanding the feature set to include customer behavioral data and social network information could refine the predictive accuracy and enable more nuanced customer segmentation.

In conclusion, this research confirms that machine learning models, when appropriately trained and tuned, serve as valuable tools for churn prediction in the telecommunications industry. The insights gained from this study pave the way for more personalized and data-driven approaches to customer retention, offering significant competitive advantage in a highly dynamic market.

VII. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who supported and guided me throughout this research project. Special thanks to my mentors and advisors for their invaluable insights and encouragement. I am also thankful to the data providers and open-source communities whose resources and tools made this study possible. Finally, I appreciate the support and motivation from my family and friends during the course of this work.

Finally, we want to sincerely thank our family and friends for their continued support, patience, and motivation in the course of this research. Their encouragement was an important source of resilience to help us power through challenges and remind us of the commitment of our endeavor.

REFERENCES

[1] Verbeke, Willem, et al. "Building comprehensible customer churn prediction models with advanced rule induction techniques." *Expert Systems with Applications*, vol. 39, no. 14, 2012, pp. 11778-11790.

[2] Huang, Bin, et al. "Customer churn prediction using improved balanced random forests." *Expert Systems with Applications*, vol. 39, no. 15, 2012, pp. 11967-11974.

[3] Nguyen, Hung, et al. "Churn prediction in mobile social games: towards a complete assessment using survival ensembles." *Data Mining and Knowledge Discovery*, vol. 29, no. 4, 2015, pp. 1065-1089.

[4] Guo, Bin, et al. "Customer churn prediction for telecom using improved ensemble learning model." *IEEE Access*, vol. 7, 2019, pp. 139434-139443.

[5] Wang, Tianshu, et al. "A hybrid model of customer churn prediction based on ensemble learning." *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 4, 2018, pp. 2063-2073.

[6] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." *R News*, vol. 2, no. 3, 2002, pp. 18-22.

[7] Chen, Tianqi, and Carlos Guestrin. "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[8] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine Learning*, vol. 20, no. 3, 1995, pp. 273-297.

[9] Verbeke, Willem, et al. "Churn prediction in the telecommunication industry: A survey of machine learning techniques." *Journal of Applied Statistics*, vol. 44, no. 10, 2017, pp. 1857-1874.

[10] Taylor, Steven J., and Richard Letham. "Forecasting at scale." *The American Statistician*, vol. 72, no. 1, 2018, pp. 37-45.

[11] Tang, Shiguang, et al. "Customer churn prediction using ensemble learning in telecom industry." *Journal of Communications*, vol. 10, no. 7, 2015, pp. 579-586.

[12] Kuhn, Max, and Kjell Johnson. "Applied Predictive Modeling." Springer, 2013.

[13] Fang, Shuai, and Jiahong Zhang. "Handling imbalanced data in telecom churn prediction using SMOTE and Random Forest." *International Journal of Computer Applications*, vol. 182, no. 40, 2019, pp. 1-6.
[14] Ali, Ahmed, and Muhammad Imran. "An effective customer churn prediction model for telecommunication using data mining techniques." *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, 2020, pp. 395-401.
[15] Brown, Ian, and Philip Mues. "An experimental comparison of classification algorithms for imbalanced credit scoring data sets." *Expert Systems with Applications*, vol. 39, no. 3, 2016, pp. 3446-3453.