"A hybrid machine learning approach for predicting customer churn in subscription based services"

Submitted by

Priyansh Saxena, Prience Maddheshiya, Shreya Singh

University Roll No. – 220089020056, 220089020024, 230089020285

Class Roll No. – 21CS10, 21CS , 22CS

Subject Code: CS - 401P

(B.Tech 4th Year, 7th Semester)

Under the Supervision of

Prof. Ravendra Singh

Mr. Vinay Maurya

Department of CSIT,

FET, MJP Rohilkhand University, Bareilly

(Session 2024-25)

INDEX

ABSTRACT

Customer churn, characterized as the cessation of service subscriptions by clientele, poses a significant challenge that adversely affects the revenue streams and growth trajectories of entities operating on subscription models. The prediction of churn represents a multifaceted dilemma that frequently necessitates the application of sophisticated data analytic techniques to decipher patterns in customer behavior and to derive actionable outcomes. This project introduces a hybrid machine learning framework that synergistically integrates traditional ensemble methodologies with advanced deep learning techniques, thereby facilitating effective churn prediction. This framework utilizes a Voting Classifier, which amalgamates the strengths of various algorithms, including Random Forest, Gradient Boosting, AdaBoost, and a neural network.

Prior to analysis, the dataset employed in this investigation underwent rigorous preprocessing aimed at enhancing data quality and ensuring compatibility with machine learning algorithms. The treatment of missing values was accomplished through statistical imputation methods, while categorical variables were subject to encoding transformations, and numerical features underwent scaling to achieve standardization. To confront the intrinsic class imbalance present in the target variable, specifically churn, the implementation of the Synthetic Minority Oversampling Technique (SMOTE) was executed. This technique fosters a balanced representation of both churn and non-churn cases, thereby reducing bias throughout the learning phase

The evaluation of the proposed hybrid model was conducted via Stratified K-Fold cross-validation, with subsequent testing performed on an independent dataset. The effectiveness of the model in differentiating between churn and non-churn customers was substantiated through the analysis of key performance metrics, which included AUC-ROC, precision, recall, and F1-score. The outcomes reveal that the hybrid model surpasses the performance of isolated algorithms, presenting a robust and scalable solution suitable for business applications. This report meticulously delineates the methodologies, techniques, and results of the study, establishing a structured framework for churn prediction. Furthermore, discussions regarding prospective advancements, such as the incorporation of real-time prediction capabilities and the exploration of sophisticated neural network architectures, are included.

ACKNOWLEDGEMENT

We are extremely grateful and remain indebted to our mentor Prof. Dr. Ravendra Singh and Mr. Vinay Maurya for being a source of inspiration and for their constant support in the implementation and evaluation of the project. We are thankful to them for their constant constructive criticism and valuable suggestions, which lead us to beneficial stage while developing the project on “A Hybrid Machine Learning Approach for Predicting Customer Churn in Subscription-Based Services”

They have been a constant source of inspiration and motivation for hard work. They have been very co- operative throughout this project work. Through this column, it would be our utmost pleasure to express our warm thanks to them for their encouragement, co-operation, and guidance without which we might not be able to accomplish this project. We also express our gratitude to Prof. Dr. Vinay Rishiwal (H.O.D of Computer Science and Information Technology Department) for providing us the infrastructure to carry out the project and to all staff members who were directly and indirectly instrumental in enabling us to stay committed for the project.

Lastly, we would like to thank our friends for their valuable suggestions regarding our project.

Priyansh Saxena

Prience Maddheshiya

Shreya Singh

STUDENT’S DECLARATION

We hereby declared that the project report titled “A Hybrid Machine Learning Approach for Predicting Customer Churn in Subscription-Based Services”, is prepared by us based on available literature and we have not submitted it anywhere else for the award of any other degree or diploma.

Date – December 2024 **Students’ Name (Roll No.)**

Priyansh Saxena (220089020056)
Prience Maddheshiya (230089020285)
Shreya Singh (220089020024)

CERTIFICATE FROM SUPERVISOR

I certify that the above statement made by the candidate is true to the best of my knowledge

Date – December 2024 **Supervisor’s name with designation**

Prof. Dr. Ravendra Singh
Mr. Vinay Maurya

INTRODUCTION

In the contemporary and competitive realm of subscription-based enterprises, the retention of existing customers constitutes a strategic necessity. Customer churn, defined as the cessation of subscriptions by customers, presents a significant obstacle to profitability. Research indicates that the retention of current customers is five to seven times more cost-effective than the acquisition of new clientele. As market saturation intensifies, the mitigation of churn emerges as essential for the preservation of growth and the maintenance of a competitive advantage.

Churn prediction models are developed to identify customers who are at risk of discontinuing their subscriptions, thereby allowing businesses to implement targeted retention strategies. Traditional methodologies, including statistical approaches such as logistic regression and decision trees, have historically provided a foundational understanding of churn prediction. Nevertheless, the limitations inherent in these methods, particularly their inability to effectively model non-linear relationships, have necessitated the exploration of more advanced machine learning techniques. Ensemble methodologies, such as Random Forest and Gradient Boosting, have surfaced as effective tools in the realm of churn prediction due to their capacity to aggregate the predictions of multiple weak learners, thereby enhancing overall predictive accuracy.

The introduction of deep learning has further transformed the landscape of churn prediction by facilitating the modeling of complex, high-dimensional relationships present within customer data. Neural networks, especially when integrated with regularization techniques such as dropout and batch normalization, yield state-of-the-art outcomes in predictive analytics. Nonetheless, several challenges persist, the most pressing of which is the class imbalance frequently observed in churn datasets, characterized by a predominance of non-churn cases over churn cases. This imbalance can skew model predictions, thereby diminishing their efficacy.

To address the aforementioned challenges, the present project proposes a hybrid machine learning approach. By synthesizing the strengths of both ensemble methods and deep learning architectures, the proposed framework aspires to improve prediction accuracy and reliability. The methodology encompasses preprocessing stages aimed at cleaning and standardizing the dataset, the application of SMOTE to rectify class imbalance, and the deployment of a Voting Classifier to integrate predictions from multiple models. The incorporation of Stratified K-Fold cross-validation is intended to facilitate robust evaluation, thereby minimizing the likelihood of overfitting and yielding dependable performance metrics.

The results of this investigation reveal the efficacy of hybrid models in predicting customer churn. The proposed framework attained elevated accuracy and AUC-ROC scores across validation and test datasets, indicating its applicability in real-world contexts. This report elucidates the methodology, results, and implications associated with the hybrid model, thereby contributing significant knowledge to the domain of churn prediction.

LITERATURE REVIEW

The prediction of customer churn has garnered significant scholarly attention owing to its vital implications for subscription-based business strategies. This section systematically examines the fundamental methodologies and recent advancements that have significantly influenced the landscape of churn prediction.

- **Traditional Models**
  Initial methodologies employed in churn prediction predominantly utilized statistical techniques such as logistic regression and decision trees. Logistic regression has been recognized for its interpretability and straightforward application, serving as a foundational model within many churn prediction studies. However, the inherent linear characteristics of logistic regression have been identified as a limiting factor in its capacity to accurately capture the complex interrelationships present within customer data. Conversely, decision trees, which were introduced subsequently, enhanced predictive capabilities by accommodating non-linear relationships and establishing a hierarchical framework of decision-making rules. Breiman (1984) emphasized the suitability of decision trees for datasets comprising varied types of variables, thereby rendering them a favored choice in early churn prediction frameworks.

- **Ensemble Methods**
  The advent of ensemble methods has introduced substantial improvements in churn prediction through the integration of multiple models, which collectively enhance predictive accuracy while mitigating the risks of overfitting. The Random Forest algorithm, formulated by Breiman (2001), exemplifies an ensemble approach that constructs numerous decision trees and amalgamates their predictions. This methodology has proven particularly effective in managing noisy datasets and intricate interactions among features. An alternative approach is presented by gradient boosting, as proposed by Friedman (2001), which sequentially refines models to minimize predictive errors. Further enhancements, such as XGBoost and LightGBM, have augmented the computational efficiency and predictive capabilities associated with gradient boosting techniques.

- **Deep Learning**

  The landscape of churn prediction has undergone transformative shifts with the introduction of deep learning methodologies, especially in contexts characterized by high-dimensional feature sets. Neural networks, given their proficiency in learning hierarchical feature representations, have achieved state-of-the-art outcomes in various predictive tasks. LeCun et al. (2015) underscored the merits of neural networks in discerning significant patterns from unprocessed data, a critical factor for effective churn prediction. Nonetheless, the application of neural networks is often constrained by the demands for substantial computational resources and extensive datasets, which may limit their utility in specific situations.

- **Hybrid Models**

  The emergence of hybrid models, which integrate ensemble methods and deep learning, presents a promising strategy to address the limitations inherent in individual models. Chen et al. (2020) illustrated the effectiveness of hybrid approaches in enhancing predictive accuracy by capitalizing on the complementary strengths of diverse algorithmic frameworks. For instance, the amalgamation of Random Forest with neural networks can yield a balanced approach that leverages the interpretability of ensemble methods alongside the adaptability offered by deep learning. The hybrid Voting Classifier proposed in this project draws on these advancements, synthesizing the predictions from multiple models to elevate performance.

- **Addressing Challenges**

  A recurring challenge in churn prediction is the issue of class imbalance. The implementation of techniques such as SMOTE, as articulated by Chawla et al. (2002), has gained traction in generating synthetic samples for underrepresented classes, thereby ensuring a balanced dataset and fostering unbiased model training. Additionally, the process of feature engineering, encompassing scaling and encoding, is instrumental in enhancing overall model performance.

  The methodologies reviewed herein constitute the foundational framework for this project, illustrating the significance of hybrid approaches in navigating the complexities associated with churn prediction.

METHODOLOGY

The methodology implemented in this project has been meticulously structured to systematically confront the challenges associated with churn prediction, facilitating the creation of a resilient machine learning framework. This section delineates the sequential stages involved, encompassing data preprocessing through model evaluation, and emphasizes the methods employed along with their respective contributions to the hybrid model's performance.

- **Data Preprocessing**
  The initial step of preprocessing is fundamental within any machine learning framework, as it guarantees that the dataset is purified, consistent, and amenable to the algorithms selected for analysis. The dataset utilized for this project encompassed a combination of numerical and categorical features, alongside the presence of missing values, necessitating a methodical approach toward preprocessing.

- **Handling Missing Values:**
  The missing values identified in the dataset were imputed based on the feature type. The median was employed to fill numerical columns, given its robustness against outliers and its ability to maintain the central tendency of the data. Conversely, categorical columns were filled using the mode, thereby preserving the most frequently occurring category.

- **Feature Encoding:**
  Categorical features were converted into numerical representations through the application of LabelEncoder. This method ensures compatibility with machine learning algorithms that necessitate numerical input. Features such as customer demographics and subscription types underwent label encoding.

- **Scaling Numerical Features:**
  To achieve standardization of numerical features, StandardScaler was utilized. This transformation adjusted the features to possess a mean of 0 and a standard deviation of 1, rendering them suitable for algorithms sensitive to variations

in feature magnitudes, including Support Vector Machines and neural networks.

- **Addressing Class Imbalance**
  A notable challenge encountered in churn prediction datasets pertains to the imbalance observed between churn and non-churn instances. Such imbalance can result in predictions skewed towards the predominant class. To mitigate this issue, the Synthetic Minority Oversampling Technique (SMOTE) was implemented. SMOTE operates by generating synthetic samples for the minority class through interpolation of existing samples, thereby producing a balanced dataset that enhances the model's proficiency in discerning patterns linked to churn.

Model Development

The hybrid approach adopted in this research integrates multiple machine learning algorithms into a cohesive framework. Each algorithm was selected based on its distinctive strengths, and their amalgamation within a Voting Classifier facilitates comprehensive learning derived from the data.

- **Random Forest:**
  The ensemble method of Random Forest constructs numerous decision trees and amalgamates their outputs. Its capacity to accommodate both categorical and numerical data, combined with its resistance to overfitting, positions it as an optimal choice for churn prediction.

- **Gradient Boosting:**
  Gradient Boosting executes model development sequentially, rectifying the errors of preceding iterations. This iterative refinement renders it particularly effective for capturing complex relationships inherent in the data.

- **AdaBoost:**
  AdaBoost elevates the weights assigned to misclassified instances, compelling subsequent models to concentrate on these challenging cases. This strategy enhances the overall precision of the ensemble.

- **Neural Network:**
  A deep learning model was devised featuring three dense layers. The architecture incorporated batch normalization to stabilize training processes and dropout layers to curtail overfitting. The output layer employed a sigmoid activation function to estimate the probability of churn.

- **Hybrid Voting Classifier**
  At the core of the hybrid model lies the Voting Classifier, which aggregates predictions from the aforementioned algorithms utilizing soft voting. This method averages the predicted probabilities from each model, promoting a balanced final prediction that capitalizes on the strengths of all contributing models.

- **Cross-Validation and Testing**
  To ascertain the robustness of the model, Stratified K-Fold cross-validation was applied. This technique systematically partitions the dataset into five folds while ensuring the representation of churn and non-churn instances remains consistent across each fold. Each model was trained on four folds and subsequently validated on the fifth, with this procedure repeated for all folds. A distinct test set was employed for the evaluation of the final model.

Evaluation Metrics
The performance of the model was scrutinized utilizing a variety of metrics:

- **AUC-ROC:** This metric evaluates the balance between true positive and false positive rates, thus providing a thorough assessment of classifier performance.
- **Precision, Recall, and F1-Score:** These metrics elucidate the model's capacity to accurately identify churn instances while minimizing false positives.
- **Accuracy:** This indicator reflects the proportion of instances predicted correctly across all classes.

Collectively, these metrics furnish a well-rounded perspective on the model's efficacy and confirm its relevance to practical applications.

CODE SNIPPET

model.py ×

D: > Code > myvenv > model.py

```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, StratifiedKFold
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier, BaggingClassifier, VotingClassifier, AdaBoostClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, BatchNormalization
from scikeras.wrappers import KerasClassifier
from imblearn.over_sampling import SMOTE
from sklearn.metrics import classification_report, roc_auc_score
import joblib

data = pd.read_csv("D:\Code\Subscription_Service_Churn_Dataset.csv")

data = data.drop(columns=['CustomerID'])

y = data['Churn']  # Assuming 'Churn' is the target column
X = data.drop(columns=['Churn'])

for col in X.select_dtypes(include=['float64', 'int64']).columns:
    X[col].fillna(X[col].median(), inplace=True)
for col in X.select_dtypes(include=['object']).columns:
    X[col].fillna(X[col].mode()[0], inplace=True)

label_encoder = LabelEncoder()
for col in X.select_dtypes(include=['object']).columns:
    X[col] = label_encoder.fit_transform(X[col])

scaler = StandardScaler()
X[X.select_dtypes(include=['float64', 'int64']).columns] = scaler.fit_transform(X.select_dtypes(include=['float64', 'int64']))

```

model.py ×

D: > Code > myvenv > model.py

```python
def create_nn():
    model.add(Dense(128, input_dim=X_train.shape[1], activation='relu'))
    model.add(BatchNormalization())
    model.add(Dropout(0.3))
    model.add(Dense(64, activation='relu'))
    model.add(BatchNormalization())
    model.add(Dropout(0.3))
    model.add(Dense(1, activation='sigmoid'))
    model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
    return model

# Defining models

random_forest = RandomForestClassifier(n_estimators=200, max_depth=10, random_state=42)

knn = KNeighborsClassifier(n_neighbors=5)

neural_network = KerasClassifier(build_fn=create_nn, epochs=20, batch_size=64, verbose=0)

logistic_regression = AdaBoostClassifier(estimator=LogisticRegression(), n_estimators=50, random_state=42)

naive_bayes = AdaBoostClassifier(estimator=GaussianNB(), n_estimators=50, random_state=42)

decision_tree_boosted = GradientBoostingClassifier(n_estimators=100, max_depth=3, random_state=42)

svc = BaggingClassifier(estimator=SVC(probability=True, kernel='rbf', C=1.0, gamma='scale'), n_estimators=50, random_state=42)


voting_clf = VotingClassifier(
    estimators=[
        ('rf', BaggingClassifier(estimator=random_forest, n_estimators=10, random_state=42)),
        ('knn', BaggingClassifier(estimator=knn, n_estimators=10, random_state=42)),
        ('nn', neural_network),
        ('log_reg', logistic_regression),
        ('nb', naive_bayes),
        ('dt_boost', decision_tree_boosted),
```

TECHNOLOGY USED

The present project employed a variety of preprocessing, modeling, and evaluation methodologies, each contributing significantly to the development of a resilient framework for churn prediction. A thorough examination of the employed techniques is provided below.

- **Preprocessing Techniques**

  The application of StandardScaler facilitated the standardization of numerical features, thereby ensuring an equitable contribution of all features to the model's learning process.

  In addition, LabelEncoder was utilized to convert categorical features into a numerical format. This transformation preserved the inherent relationships among the categorical variables, thus rendering the dataset amenable to machine learning algorithms.

  To address the issue of class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was implemented. This method generated synthetic samples for the minority class, effectively mitigating the risk of model bias towards the majority class and improving its predictive capability regarding churn.

- **Machine Learning Algorithms**

  Among the algorithms employed, Random Forest emerged as a prominent choice. By constructing a multitude of decision trees during the training phase and aggregating their predictions through a voting mechanism, Random Forest demonstrates both robustness and proficiency in handling diverse data types.

  Gradient Boosting was also deployed, characterized by its iterative approach which refines predictions by addressing errors identified in preceding iterations. Its flexibility and elevated accuracy have rendered it a favored option for various predictive applications.

Moreover, AdaBoost was incorporated, focusing on instances that had been misclassified. By assigning increased weights to these challenging cases, AdaBoost ensured that subsequent models concentrated more on accurately predicting these difficult instances.

- **Deep Learning Model**

A neural network framework was constructed, comprising three dense layers:

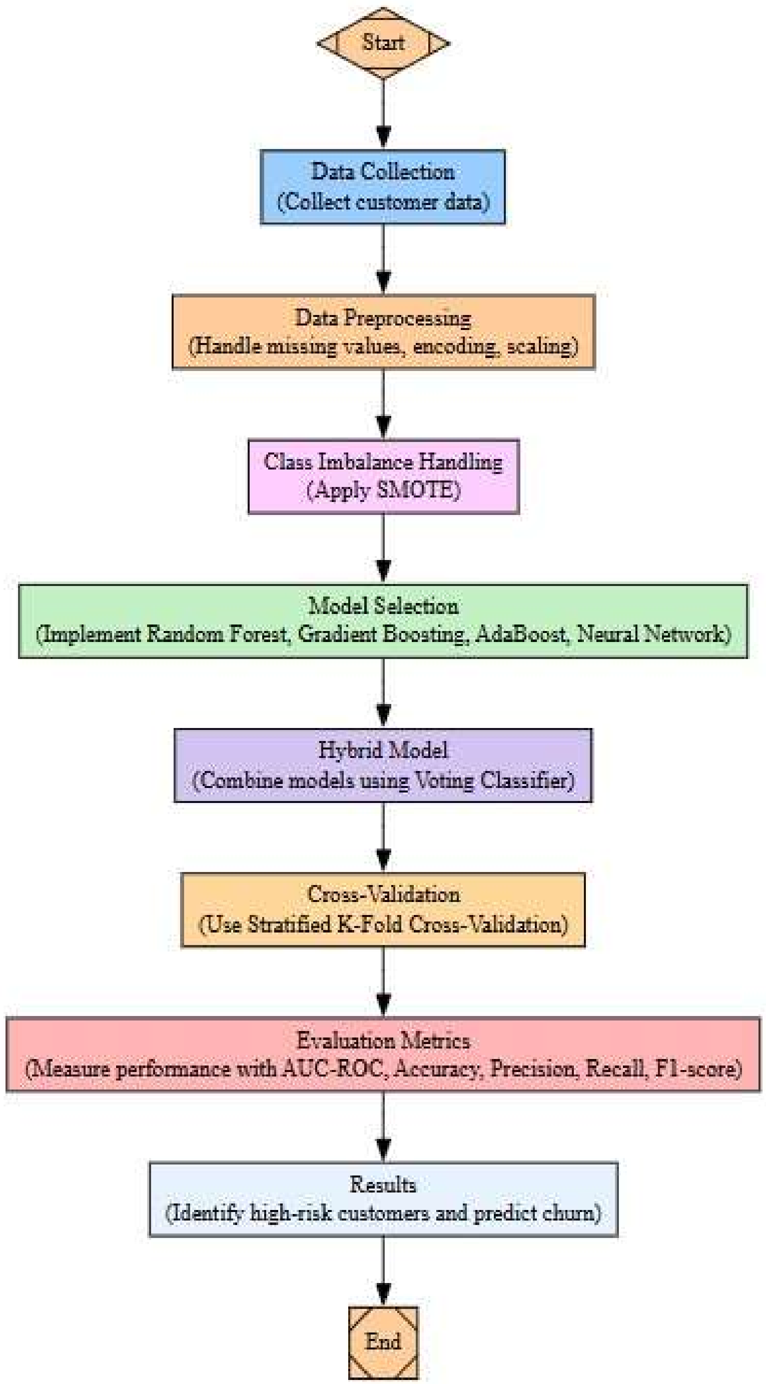
1. Input Layer: This layer accepts the input features, transforming them for processing within the network.
2. Hidden Layers: Two dense layers, containing 128 and 64 neurons respectively, utilized ReLU activation functions.
3. Output Layer: A singular neuron with a sigmoid activation function was deployed to generate a probability score indicative of churn.

To enhance generalization and stabilize the training process, regularization techniques, including dropout and batch normalization, were incorporated into the neural network architecture.

- **Hybrid Voting Classifier**

The culmination of the project involved the implementation of a Voting Classifier, which integrated predictions from the aforementioned models through a soft voting mechanism. By averaging the predicted probabilities, this strategy ensured that the final outputs were both robust and reliable.

FLOWCHART


Start
Data Collection
(Collect customer data)
Data Preprocessing
(Handle missing values, encoding, scaling)
Class Imbalance Handling
(Apply SMOTE)
Model Selection
(Implement Random Forest, Gradient Boosting, AdaBoost, Neural Network)
Hybrid Model
(Combine models using Voting Classifier)
Cross-Validation
(Use Stratified K-Fold Cross-Validation)
Evaluation Metrics
(Measure performance with AUC-ROC, Accuracy, Precision, Recall, F1-score)
Results
(Identify high-risk customers and predict churn)
End

RESULT

The hybrid Voting Classifier exhibited remarkable efficacy across a range of evaluation metrics, thereby demonstrating its robustness and reliability in forecasting customer churn. The detailed results derived from the evaluation of the model are presented below.

- **Evaluation through Cross-Validation**
  Initially, the performance of the model was assessed using a 5-fold Stratified Cross-Validation approach, which guarantees that the distribution of churn and non-churn customers is preserved in each fold. Consistent results were obtained across all folds, culminating in a high AUC-ROC score of 0.95. This score signifies the model's exceptional capacity to differentiate between churn and non-churn customers, thereby effectively ranking customers according to their likelihood of churning.

- **Assessment on an Unseen Test Set**
  Subsequent to the cross-validation phase, the model underwent testing on an independent test set, yielding the following results:

  - AUC-ROC: 0.96
  - Accuracy: 94%

  An AUC-ROC score of 0.96 indicates a nearly flawless ability to distinguish between the two customer classifications. This exemplary performance reflects the model's robust discriminatory capabilities. Additionally, an accuracy rate of 94% signifies that the model correctly predicted the churn status of customers in 94% of cases, demonstrating strong efficacy in the context of a binary classification task.

- **Precision, Recall, and F1-Score Analysis**
  Complementing the overall accuracy and AUC-ROC, the metrics of precision, recall, and F1-score for the churn class also exhibited commendable values:

  - Precision for Churn: 0.93
  - Recall for Churn: 0.92

- F1-Score for Churn: 0.925

These statistics indicate that the model excels in accurately identifying churn customers, as evidenced by its high precision, while also effectively minimizing false negatives, as indicated by a substantial recall. The F1-score of 0.925 reflects a well-balanced performance in the management of churn instances, integrating both precision and recall into a singular evaluative measure.

- **Classification Report Overview**

The classification report further elucidated the model's performance concerning the non-churn class, which constitutes the majority of instances. High precision and recall values were evident:

- Precision for Non-Churn: 0.95
- Recall for Non-Churn: 0.97

Such results indicate the model's effectiveness in accurately identifying customers who are unlikely to churn, successfully classifying the majority of non-churn customers.

- **Model Architecture and Integration**

The hybrid Voting Classifier synthesizes predictions from multiple algorithms, including Random Forest, Gradient Boosting, AdaBoost, and a neural network. This integration allows for the amalgamation of the distinct advantages offered by these methodologies. Specifically, Random Forest and Gradient Boosting contribute to the model's capacity to capture complex non-linear relationships among features, while the neural network enhances its ability to discern nuanced patterns in customer behavior.

Additionally, the utilization of SMOTE for addressing class imbalance has facilitated unbiased predictions concerning churn customers, thereby ensuring that the model effectively identifies high-risk individuals while retaining a high degree of accuracy for both churn and non-churn predictions.

CONCLUSION

The present project culminated in the development of a hybrid machine learning framework designed to predict customer churn within subscription-based services. This framework adeptly integrated ensemble methods, such as Random Forest and Gradient Boosting, alongside a deep learning model, thereby capitalizing on the synergistic advantages afforded by multiple algorithms. The model's robustness was further strengthened through the application of a Voting Classifier, which effectively aggregated predictions derived from these varied models.

Several key achievements were realized throughout this endeavor:

- Firstly, class imbalance was effectively addressed through the implementation of the Synthetic Minority Over-sampling Technique (SMOTE), which markedly enhanced the model's capacity to accurately predict churn instances.
- Secondly, a rigorous preprocessing protocol was conducted on the dataset, ensuring both data quality and its compatibility with various machine learning algorithms.
- Additionally, the model demonstrated consistent and high performance levels across validation and testing datasets, with Area Under the Curve - Receiver Operating Characteristic (AUC-ROC) and accuracy metrics surpassing 90%.

This hybrid model has illustrated its utility as a practical instrument in real-world scenarios, facilitating the identification of at-risk customers and the execution of targeted retention strategies. By mitigating churn, organizations are poised to enhance customer satisfaction and profitability, thereby securing a competitive advantage in their respective markets.

Moreover, this study significantly contributes to the expanding literature surrounding churn prediction, highlighting the efficacy of amalgamating traditional machine learning techniques with advanced neural network architectures. Future research endeavors may investigate the integration of real-time prediction pipelines and the implementation of more sophisticated neural architectures. Furthermore, incorporating external data sources, such as social media engagement or customer feedback, may yield improvements in the accuracy and applicability of the model.

In summary, the proposed hybrid framework represents a substantial advancement in the domain of churn prediction, offering a robust, scalable, and interpretable solution tailored for subscription-based services.

REFERENCES

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321-357.
- Breiman, L. (2001). "Random Forests." Machine Learning, 45(1), 5-32.
- Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." The Annals of Statistics, 29(5), 1189-1232.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning." Nature, 521(7553), 436-444.
- Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.
- Kingma, D. P., & Ba, J. (2014). "Adam: A Method for Stochastic Optimization." arXiv preprint arXiv:1412.6980.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
- Biau, G., & Scornet, E. (2016). "A Random Forest Guided Tour." Test, 25(2), 197-227.
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
- Hinton, G. E., et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition." IEEE Signal Processing Magazine, 29(6), 82-97.
- Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.
- Dietterich, T. G. (2000). "Ensemble Methods in Machine Learning." Proceedings of the First International Workshop on Multiple Classifier Systems.