National College of Ireland

Enhancing Customer Churn Prediction in the Telecom Sector Using Advance Machine Learning Techniques and Explainable AI

MSc Research Project
MSc Data Analytics

Shreyas Bhargav Bhushan
Student ID: x23175851

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shreyas Bhargav Bhushan
Student ID:	x23175851
Programme:	Msc in Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	29/01/2025
Project Title:	Enhancing Customer Churn Prediction in the Telecom Sector Using Advance Machine Learning Techniques and Explainable AI
Word Count:	8165
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th January 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keepa copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Customer Churn Prediction in the Telecom Sector Using Advance Machine Learning Techniques and Explainable AI

Shreyas Bhargav Bhushan
x23175851

Abstract

In telecommunications industry, customer churn is one of biggest issues that is faced by the operators, it is phenomenon where the customers stop using their services or switch to other operators. Telecommunications companies profitability can be heavily impacted by high churn rates, as acquiring new customers is often easy when compared to keeping existing customers. The study proposes a robust and explainable machine learning model to predict and control the customer churn. The proposed framework integrates heterogenous multi-stacking of ensemble technique, combining base models like random Forest, XGBoost, k-Nearest Neighbors (KNN) with logistic regression as meta model. To select the significant features we have applied Recursive Feature Elimination (RFE), and Synthetic Minority Oversampling Technique (SMOTE) was implemented to handle the imbalance of the class. Stratified k-fold was applied to cross validate the performance of the models. The multi-stacked model outperformed all the base models with an accuracy of 81%, while maintaining balance between recall and precision. The evaluation metrics like Accuracy, Precision, Recall, F1-score, ROC-AUC score and confusion matrix was used to validate the efficiency of the model. To address the "black box" nature of the ensemble model, the Explainable AI technique called as SHapley Additive exPlanations (SHAP) was used to improve the interpretability of the models, the technique provided insights for both global and local important features. SHAP helped to identify the significant features influencing the churn like contract type, tenure, and monthly charges. These insights help to gain the trust of the stakeholders and design targeted retention strategies.

Keywords— *Customer Churn Prediction, Random Forest, XGBoosting, Logistic Regression, k-Nearest Neighbor SHapley Additive exPlanations (SHAP).*

1 Introduction

The telecommunications industry is a significant part of the digital society (Poudel, 2024), the sector also promotes global connectivity and enables easy flow of information. Over time, the industry has changed dramatically by focusing on entire digital ecosystems rather than just voice and text as traditionally did. The sector started with a high level of monopoly and very high barriers to entry, however because of technology advancements , regularization and market near saturation the industry has become very competitive (Kirgiz, 2024). The mobile number portability (MNP) has significantly reduced the switching cost for the customers as now the customers can easily switch the providers without having to change the mobile numbers.

In this competitive world, one of the significant factors that has drawn the attention of telecom operators is customer retention. The saturation of the market and ease of customers in changing the operator has made the existing customer more values when compared to previous decade where acquiring new customers was important. Retention is crucial since the cost of getting a new customer is way higher than that of keeping an existing customer (Wagh, 2024). Customers are equally

important in any business as they keep the business going hence losing these customers is detrimental to the entities operation, for the telecom firms, customer churn is the biggest threat which leads to revenue loss and reduced market share . The increased number of customers and choices forces them to consider churn management an essential part of maintaining profits and market share as customers continue to demand options and flexibility (Poudel, 2024).

Customer churn, defined as the phenomenon of customer leaving the operator and switching to competitors due to product dissatisfaction, increase in price or poor services (Usman-Hamza, 2024). In order to address this issues, the telecom industries utilizes churn prediction models. These models learn the customer data and predict which customer may leave the operator and which factors influenced their decision. This helps the companies to take necessary measures including lowering the prices, improving the customer service or addressing customer complaints (Aggarwal, 2024) (Wagh, 2024).

As telecom industry continues to grow, the significance of machine learning model in churn prediction cannot be overstated. The study has made use of several machine learning model along with multitasking ensemble model to identify the customer churn prediction or evaluation. This study helps to state that through the application of modern and enhanced forms of ML and data analysis, telecom operators can effectively control for churn-related risks and enhance customer satisfaction hence maintaining competitiveness in the market.

1.1 Background

Telecommunications industry is one of the most important industries in the world as it offers essential communication for individuals as well as companies. However, the competition is increasing because of market saturation, the telecom companies are several serious challenges in retaining the customers. Customer churn, it is an act of customers porting or switching to another service providers which is imposing significant challenges on the telecom provides, it is also imposing financial challenge as acquiring new customer is costlier than retaining the present customer (Wagh, 2024).

The paper (Ouf, 2024) has given an approach to churn prediction in telecom industry sector by combining the XGBoost classifiers with SMOTE-ENN technique to handle the imbalance of class. It first provides primary selection of different features and then the use of ensemble model. The study states that the performance of the framework can be enhanced by combing additional optimization techniques. From this we can say that to improve the performance we can implement machine learning techniques like ensemble models. The study (Ouf, 2024) states that heterogenous multi-layer stacking, improves the predictive performances by utilizing aggregation various point of views, when compared to single models. As established by (Aggarwal, 2024), even though these algorithms have high predictive accuracy, many model like Support Vector Machine, Gradient Boosting and Random Forests, are typically called as ‘Black Box’ models. This makes the decision-makers to trust results of the models and take decisions based on the results to solve the problems.

However another difficulty for churn prediction is that the dependent variable is imbalanced, the number of non-churning customers is usually more than the number of churners (Usman-Hamza, 2024). This imbalance can affect the performance of the model significantly, which will make it difficult to identify the churners perfectly. Methods like Synthetic Minority Over-Sampling Technique (SMOTE) have been used to handle the balancing the data distribution (Aggarwal, 2024). However, the integration of SMOTE with advanced ensemble model is not explored much while trying to enhance the interpretability of the model.

Thus, the motivation of this study is to build an ensemble heterogenous multi-stacking model for increasing accuracy and the interpretability of the results gained for churn prediction. While the

SMOTE method was used for handling the imbalance of data and SHAP was implemented for interpretability, which puts the telecom operators in position to predict churn more accurately and know the reason behind those churns. Recursive Feature Elimination (RFE) methodology is used for selection of relevant features and get to get rid noise in the data. Then results of base models Random Forest, XGBoost and K-nearest neighbor is stacked as input for meta model that is Logistic Regression for improving the prediction. For better and reliable results, the performance of these model is evaluation by measuring the accuracy, precision, recall, F1-score, and area under curve with stratified K-fold cross validation. This approach will allow to implement the customer retention strategies, keep churn rates low and also lower the cost of customer acquisition and becomes best choice for telecommunication companies which aim to have a competitive edge over the companies.

1.2 Research Question

The research question of the study is How can we improve the accuracy and interpret-ability of churn prediction in telecommunication industry by implementing advance machine learning techniques, particularly a heterogeneous multi-stacking ensemble learning enhanced with SMOTE and Explainable AI (XAI)?

1.3 Research Objectives

The key research objectives are as follows:

- Application of SMOTE technique to balance the dataset especially on minority class to ensure better prediction of churn.
- Develop a heterogeneous multi-stacked ensemble learning techniques, combining the strengths of models such as Random Forest, KNN, and XGBoost (Base models) with Logistic Regression (Meta model).
- Enhance the interpretability of the model by using an Explainable AI (XAI) technique called as SHapley Additive exPlanations (SHAP).

1.4 Research Outlines

The study is divided into seven different parts. The first section provides introduction to the challenge related customer churn and customer retention. This section also provides background for employing ensemble model and explainable AI (XAI) while stating the research questions and research objectives. The section 2 discusses related work on customer churn prediction, identifying the research gaps and justifying the use of multi-stack ensemble model and SHAP. The section 3 describes the methodology of the research, data preprocessing, feature selection using RFE, balancing the data with the help of SMOTE and multi-stacking the model by combining Random Forest, XGBoost and KNN. The section 4 provides the specifications of the design for customer churn prediction. The section 5 discusses the tools required for the research and implementation of the framework. The section 6 discusses the performance of the different model and the insights provided by the application of SHAP. This section also provides the criticises the design, addresses the limitations and suggest some improvements. The section 7 discusses the contribution of the research in the field of customer churn analysis and provides the direction for future work.

2 Related Work

In this relevant work section, previous research has been analyzed to justify the validity of this study. A hybrid model and machine learning and interpretability are the two sections of related work.

2.1 Machine Learning Techniques & Hybrid modelling in Customer Churn Prediction

The aim of (Usman-Hamza, 2024) study is to solve important issues like class imbalance in datasets , while providing diverse approaches in prediction. So the paper proposes to create a framework of Heterogeneous Multi-layer Stacking Ensemble method (HMSE) to enhance the Customer Churn Prediction (CCP) in telecom sector. To do so, the authors have integrated five different machine learning models, including Random Forest, Bayesian Network, Support Vector Machine, K-Nearest Neighbor, and RIPPER classifiers altogether in a multi-layer stacking ensemble model. This approach is further improved by implementing Forest Penalizing Attribute (FPA) model as a meta-classifier along side the Synthetic Minority Oversampling Technique (SMOTE) for addressing the issue of class imbalance. The study evaluates the performance of HMSE and and SMOTE enhanced version (S-HMSE) using two benchmark datasets based on the accuracy, Area Under Curve (ACC), F-measure and Matthews Correlation Coefficient (MCC). The study demonstrates that the proposed S-HMSE model outperforms all other baseline classifiers and provides a significant increase in AUC and MCC compared to homogenous ensembles and present Customer Churn solutions. The research analyzes the problem of class imbalance within the CCP while identifying the drawbacks of a high computational cost and less diverse application to other domains. To improve the model, future research proposed using sophisticated feature engineering methods, expanding the samples with different sets of data and using mixed hybrid approaches to minimize computational expenses. In essence, the current study advances the knowledge of CCP meaningfully and provides insights helpful to practical application of decision analytical tools for designing customer retention strategies within the telecommunication industry.

The (Wagh, 2024) research aims to develop a very efficient machine learning models to solve the customer churn prediction problems in telecom. In the study the data is prepared by excluding all unwanted and incomplete data and by converting the categorical variables into numerical variables. In order to tackle the question of class imbalance, the study uses Synthetic Minority Oversampling Technique (SMOTE) together with Edited Nearest Neighbor (ENN), which helps to bring better balance of non-churn and churn classes. In the paper, the feature selection is performed on the basis of Pearson Correlation formula to determine the significant feature influencing the churn, like contract type, tenure and monthly charges. The Random Forest model displayed better performance when compared to decision tree after balancing the data. With the help of Cox Proportional Hazard model and survival analysis to predict the timing and probability of churn for customers, which provides the companies with necessary information to design customer retention strategies.

The (Ouf, 2024) research focused on overcoming the issues like class imbalance and data quality by developing a mixed approach for enhancing the customer churn predictions in telecom sector. The study proposes a framework combining three layers, namely data preprocessing, algorithm classification and evaluations. In the first layer, the feature engineering is performed by scaling through standardization and selecting the important features by employing univariate analysis and PCA is combined with SMOTE-ENN resampling for feature reduction. In the second for classification, the study has employed XGBoost algorithm which is suitable of handling high dimensional and imbalanced data. The third layer, evaluation layer accuracy, precision, recall, F1-score, and AUC-ROC are considered to evaluate the performance of the model before and after

balancing the data. The proposed model achieved an impressive accuracy of 99.2% by surpassing the existing methods. The study helps to solve the previous research question of integration of ensemble model and resampling techniques. However, the limitations are like excess computational time which can be overcome by implementing hybrid ensemble models.

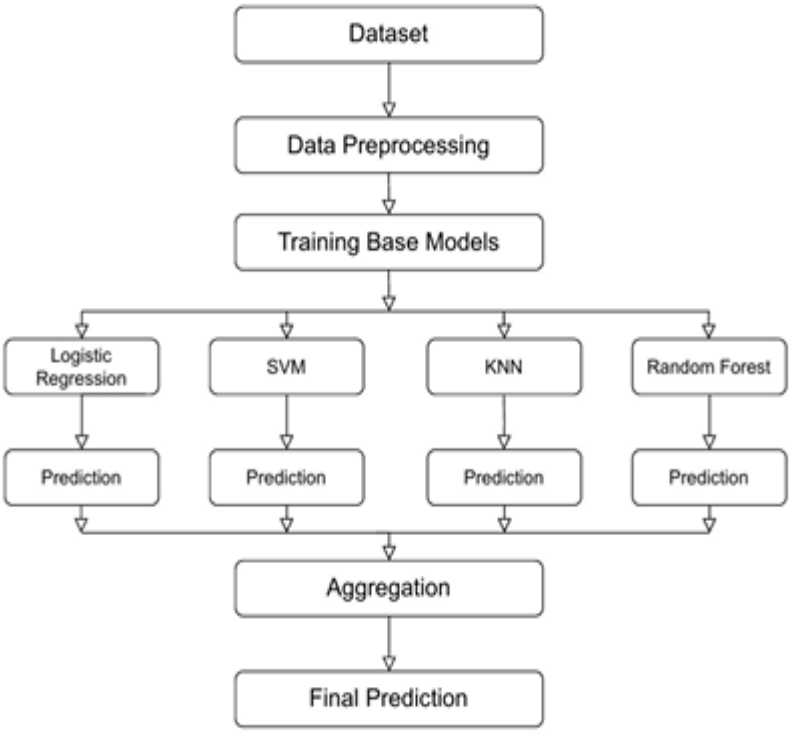

Fig.1. Architecture of Hybrid Churn Prediction by Pallav Aggarwal (2024)

The _(Aggarwal, 2024), presents an advanced Hybrid Churn Prediction (HCP) model by combining multiple machine learning models to improve the churn predictions of the telecom industries. Advanced ensemble techniques like Random Forest and Gradient Boosting have improved the performance by increasing accuracy and robustness. However, their black-box nature limits their interpretability, which makes it difficult for the stakeholders to trust. The Hybrid Churn Prediction (HCP) involves Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines (SVM), k-nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP). These models are used in HCP framework with an ensemble mechanism that integrates high accuracy of pattern identification for each model. The study states that when compared to the present models of machine learning and deep learning the HCP model presented by the research displayed better performance. However, there is need for improvement in real-time predictive capabilities and comprehensive feature engineering, limiting practical solutions.

The (Kavitha, 2024), proposes combing supervised and semi-supervised learning technique to improve customer churn prediction model reliability. In the research the methodology involves, data preprocessing, feature engineering employing Random Forest and Gradient Boosting to identify the key predictors and supervised models like RF, KNN, SVM and Gradient Boosting. The study make use of pseudo-labeling to address to address the challenges encountered by limited labelled datasets which is applied to assigned predicted labels of the data to be labeled, therefore improving the quality of dataset. Then the research has made use of semi-supervised algorithms for further predictions including Self-Training and Label Propagation. The study shows that for both supervised and semi-

supervised classifications, the Random Forest and Gradient Boosting produces the highest accuracies with pseudo-labelling significantly improving the generalization. Also, the effectiveness of iterative improvements is shown by the performance of Label Propagation by self-training. However, the study suggests pseudo-labelling is exposable to noises and also that semi-supervised model has high computational complexities.

2.2 Interpretability in Customer Churn Prediction

The _(Wang, 2024) study aims to increase the performance of churn prediction by suggesting a hybrid framework that combines the Fully Connected Layers Convolutional Neural Networks (FCLCNN) and Long Short-Term Memory (LSTM), however it seems to have issues with interpretability particularly in the aspect of designing customer retention strategy. The research tries to solve the problem by providing information to contribution of features and plan to decision making by implementing methods like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). The localized explanation given by LIME helps to handle customer with a personalised approached whereas SHAP helps to target particular customers with retention strategies by proving the feature importance scores. Despite these advantages, SHAP is computationally very expensive and LIME is relatively unstable for extremely large dataset. However, this can be addressed by simplifying the architecture to prioritize the key feature identified feature selection methods which could reduce the computational cost while maintaining the performance.

Similarly, this _(Poudel, 2024) study aims to enhance the accuracy and interpretability of customer churn prediction by combining SHAP with machine learning models. The SHAP helps to develop personalised strategies by identifying the significant features, and the local explanations give insights about individuals who are likely to churn. However, the study suggests that SHAP has computational problem related to scalability and it difficult to interpret without the knowledge of the domain. The study also argues that model tend to focus on accuracy of the predictions rather than their interpretability is not useful for real world problems.

The study _(Faraji Googerdchi, 2024), proposes to optimize the classification accuracy while enhancing the customer churn prediction by employing multi-objective evolutionary clustering (MOEEC) model. The study argues that stakeholders find it difficult to interpret the feature influencing the churn with this method, this drawback of interpretability decrease their chance in real world application as telecom industries need better interpretable model to design better retention strategies. However, this can be solved by prioritizing complexity reduction and enhancement of interpretability.

Table 1. Literature Survey

Sl. No.	Paper Name	Authors Name	Dataset	Models Used	Results Summary
1	Assessing the effectiveness of OTT services, branded apps, and gamified loyalty giveaways on mobile customer churn in the telecom industry: A machine-learning approach	Omer Bugra Kirgiz, Meltem Kiygi-Calli, Sendi Cagliyor, Maryam El Oraiby	Telecom dataset from a Turkish operator	Logistic Regression, Random Forest	Random Forest achieved 96% accuracy and 0.99 AUC. Key predictors that were identified was months in contract, gamified loyalty participation, age, call center interactions, and app logins whereas OTT services had minimal impact.
2	Telecom Churn Prediction Using Python	Aditi Sharma, Mayank Tyagi, Sunil Kumar	Dataset fromNigerian telecom company	Convolutional Neural Network (CNN), Multi-Layer Perceptron (MLP)	MLP models achieved 80% and 81%, however CNN model achieved an accuracy of 89%.
3	A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry	Shimaa Ouf, Kholoud T. Mahmoud, Manal A. Abdel-Fattah	IBM Telco Churn,Orange Telecom,Iranian Telecom	XGBoost with SMOTE-ENN	By using SMOTE-ENN resampling the hybrid framework achieved the highest accuracy of 99.92% on the Iranian dataset, 98.25% accuracy on The Orange dataset reached, and IBM Telco achieved 98%.
4	Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models	Victor Chang, Karl Hall, Qianwen Ariel Xu, Folakemi Ololade Amao, Meghana Ashok Ganatra, Vladlena Benson	Maven Analytics Telecom Customer Churn dataset	Logistic Regression, KNN, Naïve Bayes, Decision Tree, Random Forest	Random Forest achieved the highest accuracy (86.94%), AUC score (0.95), SHAP and LIME were used for to enhance the explainability.
5	An Efficient Churn Prediction Model Using ML Supervised and Semi-Supervised Learning Techniques	Ch Kavitha, Vadada Yamuna, et al.	Kaggle Churn Dataset	Random Forest, Gradient Boosting, KNN, SVM, Self-Training, Label Propagation	Supervised models achieved accuracies of 95%-96%, with Self-Training improving to 96%. Label Propagation achieved lower accuracies of 84%-85%.
6	Customer Churn Prediction in the Telecom Sector	Pallav Aggarwal, Vaidehi Vijayakumar	Customer behavior & Telecom partner data	Logistic Regression, Random Forest, SVM, KNN, Gradient Boosting, Multi-Layer Perceptron, Hybrid Churn Prediction (HCP)	HCP achieved the best results with accuracies of 79.6% (Dataset 1) and 86% (Dataset 2), outperforming individual models in precision, recall, and F1-scores, particularly for the churn class.

7	Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach	Mohammed Affan Shaikhsurab, Pramod Magadum	IBM Telco , Churn-in-Telecom Dataset, Orange Telecom Dataset	Adaptive Stacking Ensemble where the Base Models are XGBoost, LightGBM, LSTM, MLP, SVM and Meta-Learner is Logistic Regression	Achieved an accuracy of 99.89% on the Orange Telecom dataset, significantly performing better than individual models and other techniques in churn prediction.
8	Sampling-based Novel Heterogeneous Multi-layer Stacking Ensemble Method for Telecom Customer Churn Prediction	Usman-Hamza, F.E., Balogun, A.O., Amosa, R.T., Capretz, L.F., Mojeed, H.A., Salihu, S.A., Akintola, A.G., Mabayoje, M.A.	Kaggle (Dataset A) and UCI (Dataset B)	Heterogeneous multi-layer stacking ensemble (HMSE) integrating Random Forest, Bayesian Network, SVM, KNN, and RIPPER classifiers; SMOTE for class imbalance correction; Forest Penalizing Attribute (FPA) meta-model	HMSE achieved high performance with accuracy of 95.02% on Dataset A and 90.80% on Dataset B. S-HMSE, which integrates SMOTE, further improved the performance to 97.24% and 95.76%, respectively. Enhanced robustness and interpretability in handling class imbalance were noted.
9	Customer churn prediction in telecom sector using machine learning techniques	Sharmila K. Wagh, Aishwarya A. Andhale, Kishor S. Wagh, Jayshree R. Pansare, Sarita P. Ambadekar, S.H. Gawande	Telco-Customer-Churn dataset	Decision Tree, Random Forest	The Random Forest model achieved 99.09% accuracy, precision, and recall after addressing data imbalance using SMOTE and ENN. Decision Tree showed higher result after performing up-sampling. Monthly charges, tenure, and contract type are the features identified as key predictors by feature selection.
10	Explaining Customer Churn Prediction in Telecom Industry Using Tabular Machine Learning Models	Sumana S. Poudel, Suresh Pokharel, Mohan Timilsina	Kaggle Dataset	BM, Random Forest, SVC, Logistic Regression, XGBoost, Neural Networks, AdaBoost	GBM scored highest accuracy of 81% accuracy compared to other model. Contract, Monthly Charges, and Tenure as key features driving churn predictions that is highlighted by SHAP analysis.

3. Methodology

This study proposes a framework by integrating advance machine learning models with an aim to develop an accurate and reliable model for customer churn prediction. This framework is further divided into crucial steps so that the process of prediction is not only efficient but reliable and easy to implement at a large scale. It is significant to build a churn prediction system that helps to accurately predict the customers that are likely to churn based on the past data and uncover the reason, to guide prevention or reduction of churn. The framework is specifically designed to face challenges by improving the interpretability of churn predictions in telecom industry by employing advance machine learning techniques particularly a heterogenous multi-stacking ensemble models enhanced with SMOTE. By handling these challenges, the study wants to decrease the amount of customers who leave the company and increase the use of churn retention by using the right data. Additionally, it also helps to increase the interpretability of the model to make clear for the stakeholders explaining

the reason behind the predictions. The performance of the model is examined using the key evaluation metrics, in order to confirm its validity, scalability and it efficiency in predicting the customer churn and help design customer retention strategy.

3.1 Data Layer

3.1.1 Data Collection

In this research, we have made use of dataset that is publicly available on Kaggle known as IBM Telco Customer Churn Dataset. This dataset provides basis for the customer churn prediction as contains significant details about the customers. This datset is made up of 7043 rows and 21 features, where every single row is associated with an unique customer and each column containing the different features is connected to customer churn. There attributes are categorised into three primary groups, namely service related, account related, and demographic details. The service-related attributes are phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies. Customer Account related features are contract type, payment method, paperless billing status, monthly charges, and total charges. Demographic associated attributes are made up of gender, age, marital status, and dependents. The dependent variable is called "Churn" that defines whether a customer abandoned the service provider in the last months, and it is the predicted variable. This dataset is well-suited for churn prediction analysis as it offers multiple features and well-defined binary outcome. Additionally, it also supports several type of analysis including feature importance evaluation and customer segmentation. The study will develop highly reliable and easy to interpret model to predict the customer churn, to understand the reason behind customer churn and to design strategies for customer retention. The data was obtained from Kaggle website and could be downloaded completely free and used for research after agreeing with the terms of use, hence following the acceptable standard of ethics where open datasets are involved.

3.1.2 Exploratory Data Analysis

The exploration of dataset in Fig.2 reveals the features influencing the churn rates among telecom customers. The gender distribution is equal for both males and females, around 3500, however the monthly chur rate higher for males (31.43%) when compared to females (28.57%). On interesting observation is customers with month-to-month contract (45%) tend churn more while the customers with one year (12%) and two year (6%) contracts show lower churn rates, displaying that long-term contract minimize the customer churn.

When it comes payment method, electronic check displays the highest churn of 45% among 1200 non-churn and 1000 churners, on the other hand Bank Transfer and Credit card users have lower churn rates comparatively nearly 14% and mailed check users somewhere in between them. Analysis on Internet services displays that fibre optics has highest churn rates about 40%, while DSL users is 17% and the non-internet users have minimal churn rates.

Customers with partners tend to have lesser churn rates (20%) when compared to customers with partners (32%). The same is true for non-senior citizens as they churn less than senior citizens. Customers with no online security show higher churn rates of 43% whereas customers with tech support display churn rate of only 12%.

The customers who choose to paperless billing (38%) churn more than the customers who choose non-paperless billing (13%). Almost 25% of the customer with phone service tend to churn, whereas only 17% of the customer without phone service churn. Monthly charges reveal highest churn density of $70–$90 and the non-churners are at $20–$25. It can also observe that total charge churned people

tend spend most at $500–$1,000 compared to non-churned people who spend most at $1,000–$2,000, revealing that total charges have longer retention in higher values.

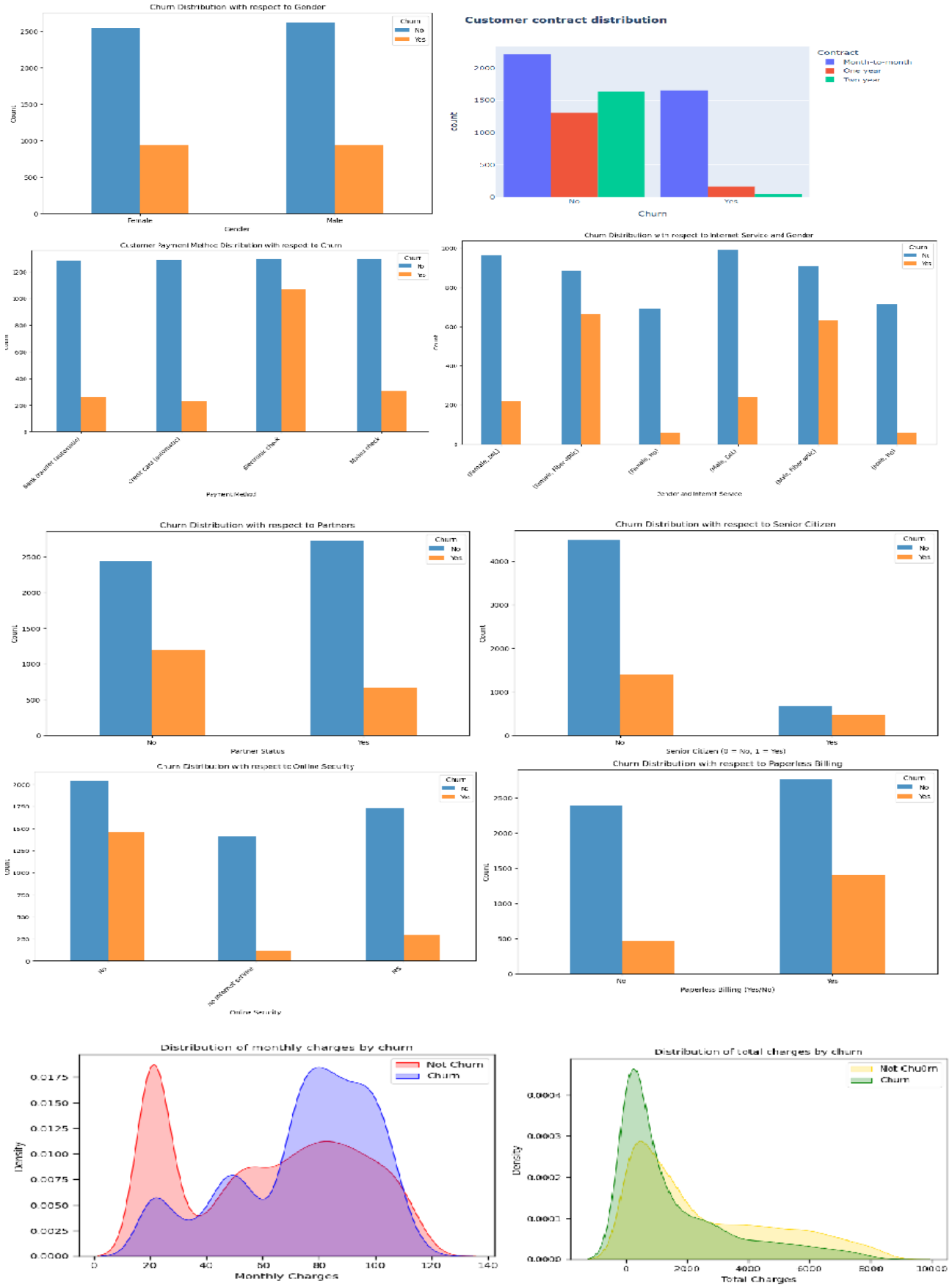

Fig.2. Data Exploration

3.1.3 Data Preparation

Data pre-processing is on of the important steps as it deals with missing values, outliers, noises and encoding of categorical variable preparing the data for feature engineering, handling class imbalances, also for multi-stacking the model, by enhancing model accuracy and interpretability for customer churn predictions.

To select the significant feature have made use of Recursive Feature Selection (RFE), as it iteratively trains the model by removing the least important attributes until the only most significant features remains. The efficiency of the model and it ability to identify the influencing attribute is enhanced by the implementation of RFE with machine learning techniques (Huijie Ji, 2022).

Class imbalances affect the customer churn prediction effectively by skewing model learning process, as the non-churners (majority class) dominates the training data. SMOTE (Synthetic Minority Over-sampling Technique) solves the class imbalance by generating synthetic samples for the minority class (the churners) in between the existing samples. It essential for the customer churn prediction as it helps the model to learn from positives and negatives and more importantly reduces the preference of the majority of the class (Shaikhsurab, 2024).

3.2 Algorithm layer

3.2.1 Data Modelling

The proposed multi-stacked model is combination of individual base models to enhance the predictive performance. Random Forest is a tree-based machine learning technique which develops several decision trees during the training and together combines the results to enhance the accuracy and reduce the problem of overfitting. A very strong predictive performance is displayed by Random Forest model as a part of multi-stacked ensemble model (Usman-Hamza, 2024). The effective prediction of customer churn in large dataset with complex customer behaviours is possible due to the ability of the Random Forest to combine the insights from multiple decision tree.

K-Nearest Neighbor (KNN) is used for regression and classification of a particular object as it is a type of non-parametric method. This process categorizes data points according to how close they are to ‘k’ nearest neighbours determined using distance measures such as distance or Euclidean distance. Its nature and instance based learning make it suitable for identifying the churn patterns in ensemble techniques for churn prediction by improving the performance (Usman-Hamza, 2024).

XGBoosting (Extreme Gradient Boosting) is a type of gradient boosting technique, it effectively handles large amount of data efficiently and accurately as it makes use of regularization and parallel computations to handle sparse data. XGBoosting highly suitable for churn prediction as it provides the understanding of the relationship between customer’s feature and churn rate, while handling large datasets and class imbalance efficiently. (Ouf, 2024).

Logistic Regression is one of the suitable techniques for customer churn prediction due to its ability to interpret the results that helps the stakeholders understand the reason behind the customers leaving the service and identify specific cases like churn and non-churn. Logistic Regression is suitable as meta model in multi-stacked model as it provides the final results with the best predictions from the base learners in the process by weighting the probabilities to inputs optimally. This increases the accuracy and reliability of the ensemble model since the algorithms involved are unique in their own way.

3.3 Evaluation and Interpretation Layer

3.3.1 Evaluation

The performance of the model is tested by using measures like accuracy, precision, recall, F1-score, and confusion matrix and AUC-ROC score. The overall correctness can be understood by the accuracy, while the churners can be identified with help of recall and precision. F1 combines the results of precision and recall. The AUC-Roc curve helps identify the models ability to classify churners and non-churners, where higher AUC values means better performance and the confusion matrix provides a detail information of true/false positives and negatives, helping to identify attributes like missed churners which help to design better retention strategies.

3.3.2 Interpretation

To improve the interpretability of the model, we made use of SHapley Additive exPlanations (SHAP) an Explainable AI technique. This technique provides better interpretation of the machine learning model by finding out the attributes that is influencing the overall predictions. When it comes to customer churn predictions. SHAP helps to gain the trust of the stakeholders and at same time helps decision-makers comprehend the model results as it focuses on important attributes affecting the churn, which can be easily understandable (Poudel, 2024).

4 Design Specification

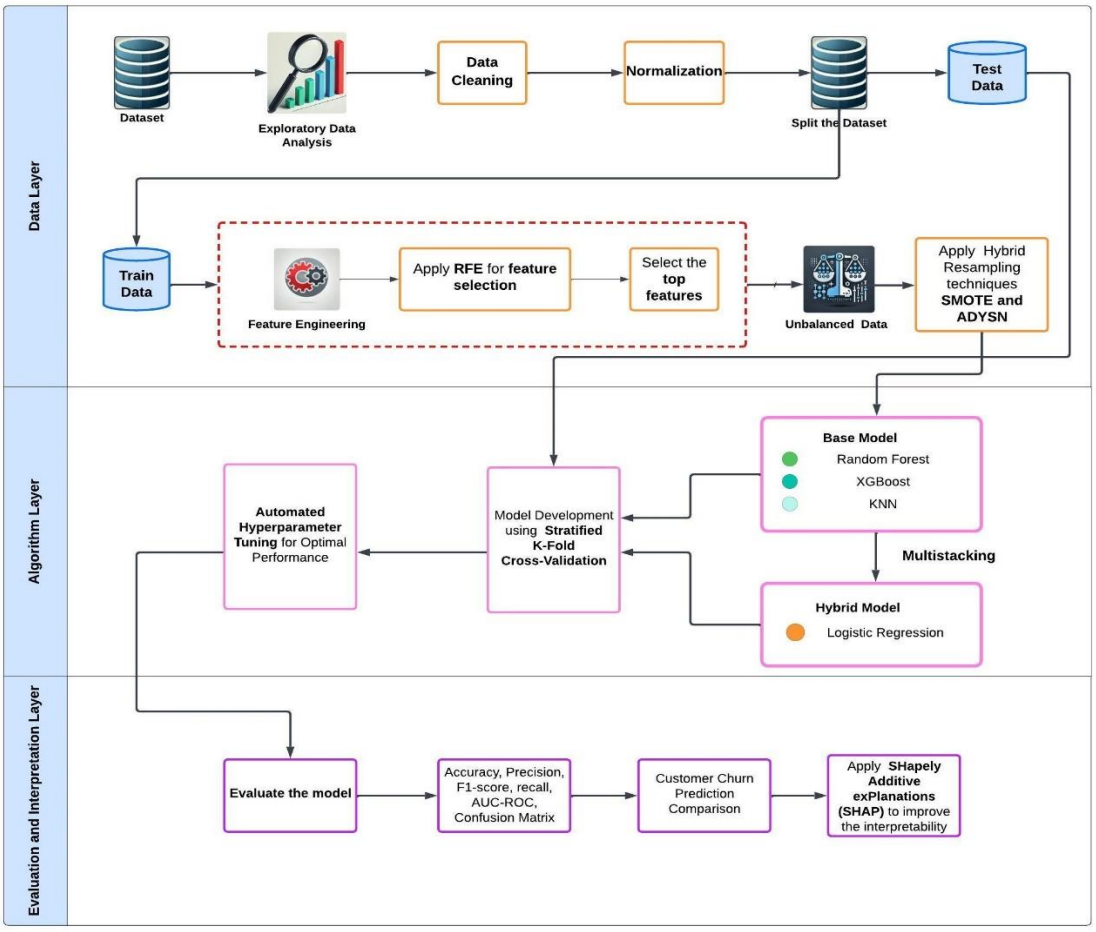

Fig.3 Design Architecture

The project proposes a framework in Fig.3, for customer churn predictions by combing state-of-art machine learning models to develop robust but interpretable model. In the data layer, the dataset is explored to understand different feature influencing the churn rates, then the missing values, outliers and noises are handled, and Recursive Feature Selection (RFE) is used for feature selection. Synthetic Minority Over-sampling Techniques is implemented to create to synthetic samples, to help model understand about churners and non-churners, while balancing the class.

In algorithm layer, where the results of the multiple base models, such as Random Forest, KNN, and XGBoost are stacked on the meta model, Logistic regression, to enhance the result of the model is accurate. Stratified k-fold cross validation is performed on the model to ensure that the proposed model is reliable, robust and generalizable across different customer behaviours and hyper-tuning of the model the performed to improve the performance.

In evaluation and interpretability layer, the performance of the model is evaluated by employing various evaluation criteria like accuracy, precision, recall, F1-score, AUC-ROC score and Confusion matrix. Lastly, SHapley Additive exPlanations (SHAP) is used to understand the individual features more efficiently, which helps to earn the trust of stakeholders and design more reliable retention strategies.

5 Implementation

The implementation of the project was performed using Python language in Juypter notebook, an environment used for iterative testing and debugging having powerful libraries and tools peculiar to machine learning. Numpy was used to manipulate the data and create feature, while pandas was used to data input and data preparation. Whereas, scikit-learn package was used for transformation of feature, model creation and evaluation of the performance. The class imbalance was addressed with help of Synthetic Minority-Sampling Technique (SMOTE) by using library called as imbalanced-learn. StandardScaler from scikit-learn library was used standardize the numerical features. This standardization made sure that numerical variables had a mean of 0 and a standard deviation of 1, to make it useful of distance based and gradient based algorithms. Ordinal variable was label encoded, while the two categorical were transformed into two binary dummy variables using one-hot encoding.

The implementation of Random Forest model was optimized by using GridSearchCv that provides a systematic approach for the hyper-tuning of the model. The technique involved specifying an extended parameters space for searching of optimal hyperparameters like number of trees (n_estimators), tree depth (max_depth) and the minimum samples needed for the terminal nodes (min_sample_split) and in leaves (min_sample_leaf). The stratified k-fold cross (5 fold) validation was employed to ensure that model evaluated effectively. SHAP values was calculated using TreeExplainer, which was used to identify the importance of the feature. Additionally, SHAP force plot was used to understand the influence of individual feature for a specific prediction. For XGBoost model we made use of RandomizedSearchCV to evaluate hyperparameters likes n_estimators (number of trees), learning_rate (step size), max_depth (tree depth), gamma (regularization). For accuracy 5-fold cross validation was used, and the search was performed over 100 randomly generated hybridizations. The classification report, confusion matrix and AUC-ROC curve was used to evaluate the best XGBoost model (best_xgb_model). Similarly to Random forest model, TreeExplainer and SHAP force plot was use to enhance the interpretability. In the same way. KNN also used RandomizedSearchCV with hyperparameters that included n_neighbors, distance metrics (minkowski, euclidean, manhattan), weight schemes (uniform, distance), and search algorithms (auto, ball_tree, kd_tree, brute) and the best parameter was identified after evaluated 100 random

combinations. Then the performance was tested using evaluation metrics. For interpretability, to calculated the feature importance of non-tree based model, SHAP made use of KernelExplainer and model's local explanation for individual predictions was given with the help of SHAP force plot.

The multi-stacked ensemble model was developed by using base models like Random Forest, XGBoost and KNN with Logistic Regression as the stacking classifier. The RandomizedSearchCV was employed for the base models to tune hyperparameters like estimators for Random Forest, learning rate and subsampling for XGBoost and number of neighbour for KNN and the meta model that is logistic regression is fine-tuned with GridSearchCV. The base models are stacked for the development of metal model by integrating the optimised models into StackingClassifier. The decision making of the model was explain with the help of SHAP. To access to predict_proba function StackedModelWrapper was created. Mean SHAP values was computed to rank features to know the global feature importance visualization and local explanations of individual predictions was given by force plot.

6 Evaluation

6.1 Experiment 1: Random Forest

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.79	0.83	1549
1	0.55	0.70	0.61	561
accuracy			0.76	2110
macro avg	0.71	0.74	0.72	2110
weighted avg	0.79	0.76	0.77	2110

Fig.4 Classification Report of Random Forest

The Fig.4 depicts the performance of the Random Forest model and provides an understanding on efficiency of model with respect to predictive performance of classification task. The classification report displays a precision of 88% of No churn which is the major class while precision of 55% for Churn that is minor class and an accuracy of 76%. The recall shows that to identify the churn features correctly the model performs at 70% for the minority class, while the F1-score for minority class is 61% displaying that recall and precision is inversely proportional to each other.

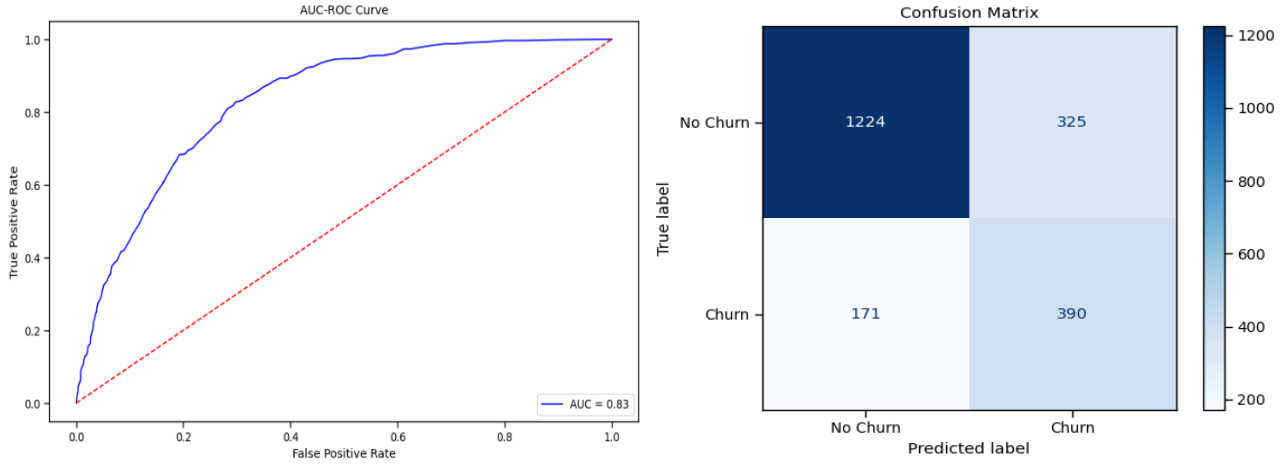

Fig.5 ROC-AUC curve and Confusion Matrix of Random Forest Model

In Fig.5 , the ROC-AUC displays the area under the curve (AUC) of 0.83, showing how accurately the Churn and No Churn classes is classified. The curve shows the ability of the model to achieve high true positives while having low false positive rate. The efficiency of the model in handling classification problems can easily be understood by this score.This score helps understand the overall efficiency of the model in handling classification issues.

The confusion matrix shows that 1224 true negatives (No Churn correctly classified) and 390 true positives (Churn correctly classified). However the confusion matrix displayed 325 false positives that is No Churn misclassified as Churn) and 171 false negatives (Churn misclassified as No churn).

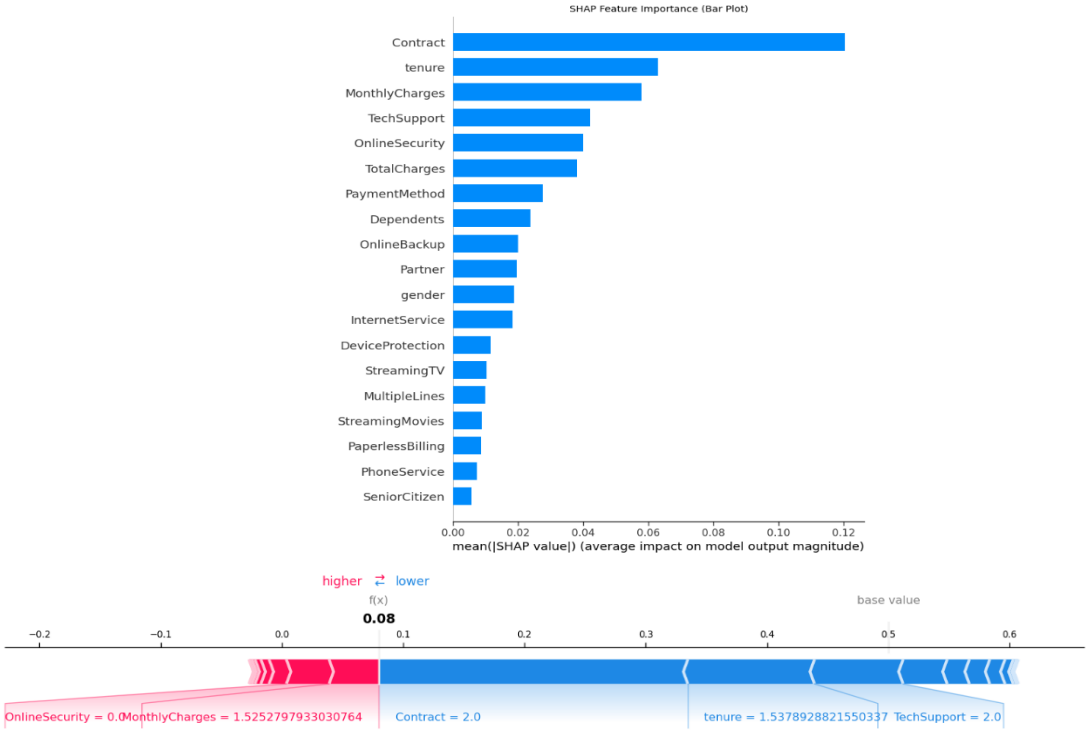

Fig.6 SHAP analysis of Random Forest

The Fig.5 reveals that key features such as Contract, Tenure and Monthly charges plays vital rule to understand the behaviour of the customers. By improving the interpretability, the Global and Local explanations display influence of individual features on the prediction of churn. To retain the customers who are the risk of churning, this transparency helps the companies to assist to enhance contracts, tenure incentives and tech support, by addressing these features the telecom companies can improve customer satisfaction, revenue loss and also implement effective customer retention strategies.

6.2 Experiment 2: XGBoost

The Fig.7 shows that the performance of the XGBoost model was evaluated using the classification report with parameters such as accuracy, precision, recall and F1-score. The model showed an accuracy of 75% with high accuracy for majority class that is No churn, it also displayed a precision of 86%, recall of 79% and F1-score of 82%, meaning the model perfectly classified who are not going to churn. However, in the case of majority class (Churn), the model displayed a precision of 53%, recall of 66%, and F1-score of 58% displaying average performance in identifying the customers at risk of leaving the service provider. The Fig.8 shows that the model with ROC-AUC score of 0.81indicates the high ability of the model to classify Churn and No Churn.

Classification Report for Tuned XGBoost:

	precision	recall	f1-score	support
0	0.86	0.79	0.82	1549
1	0.53	0.66	0.58	561
accuracy			0.75	2110
macro avg	0.70	0.72	0.70	2110
weighted avg	0.77	0.75	0.76	2110

Fig.7 Classification Report of XGBoost

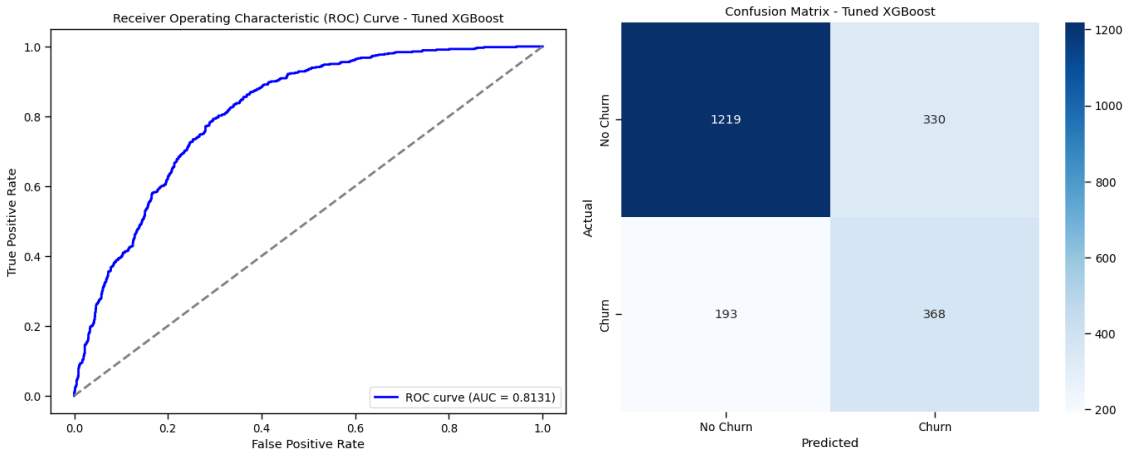

Fig.8 ROC-AUC curve and Confusion Matrix for XGBoost model

The ability of the model to classify is also displayed with the help of confusion matrix. The model correctly classified 1219 No churn (true negatives) and 368 Churn (true positives). However, the model classified 330 false positive and 193 false negatives, although then false negatives is less in number this need to be improved, as incorrectly classified churns mean missed chances to apply retention strategies.

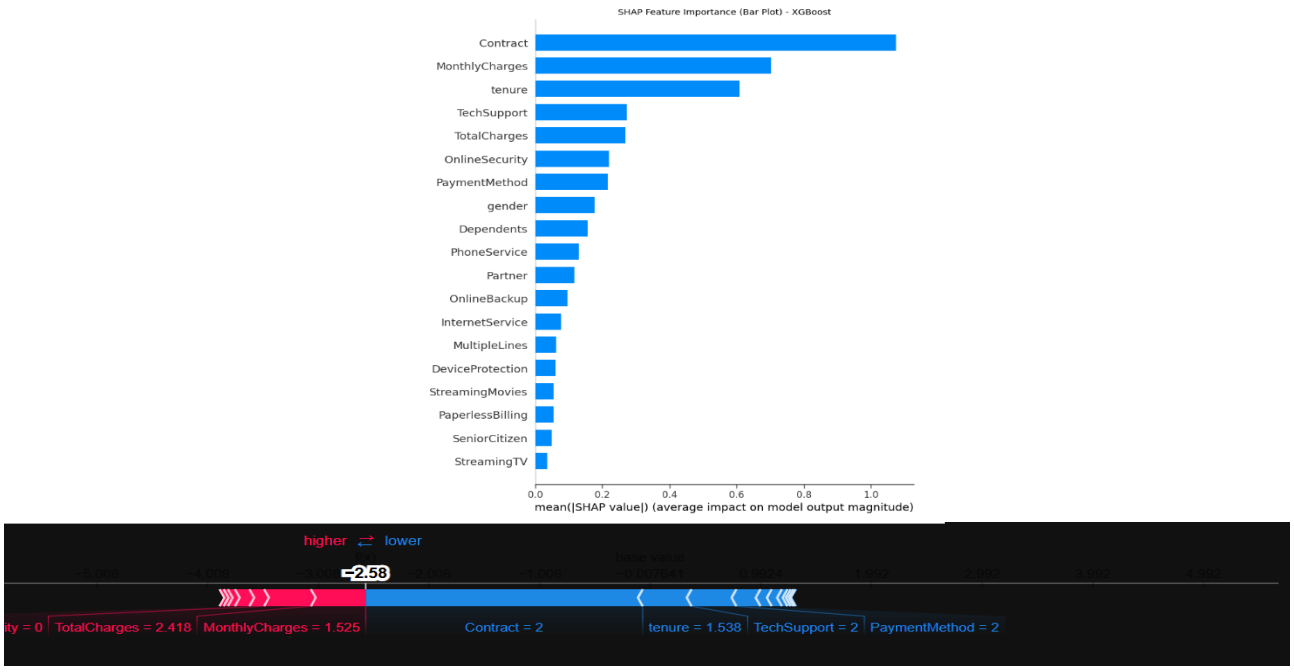

Fig.9 SHAP analysis of XGBoost

Like Random Forest model's SHAP analysis, the Fig.9 shows the SHAP analysis of XGBoost shows that features like Monthly Charges, Tenure and Contract are the key influencers for the churn, with financial factors and support services having vital role. The insights given by Loacal explanation on how TotalCharges increase the churn rate while Contract and TechSupport can help to reduc the churn rate can be used to develop targeted retention strategies to reduce the churn

6.3 Experiment 3: k-nearest neighbors (KNN)

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.86	0.86	1549
1	0.60	0.58	0.59	561
accuracy			0.79	2110
macro avg	0.73	0.72	0.72	2110
weighted avg	0.78	0.79	0.79	2110

Fig.10 Classification Report of KNN model

The evaluation of KNN model in Fig.10 shows very high performance of the model over majority of class that is the No churn with a high accuracy of 79%, which is greater than both the previous base models. The classification report displays that the model efficiently predict the customers staying with a precision of 85%, recall of 85% and F1-score of 86% for the No Churn class, while the model performs with a precision of 60%, recall of 58% and an F1-score of 59% for Churn class, identifying the customer who are likely to churn.

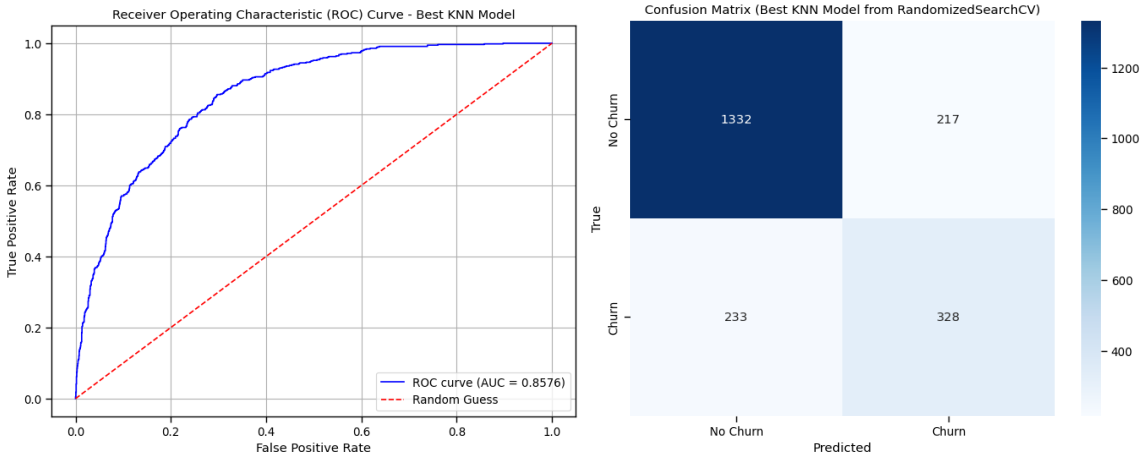

Fig.11 ROC-AUCcurve and Confusion Matrix for KNN model

The ROC-AUC score in Fig.11, of 0.85 for KNN model is also higher than the other two base models, displaying its ability to properly classify Churn and No Churn and it has also shown good balance between specificity and sensitivity.

The additional information of prediction by developed model is provided by confusion matrix, it displays that 1332 of the No Churn class and the 328 of Churn class were classified correctly by the

model. However, the model classified 217 No Churn customers and 233 Churn customers incorrectly, meaning that model still has room to improve.

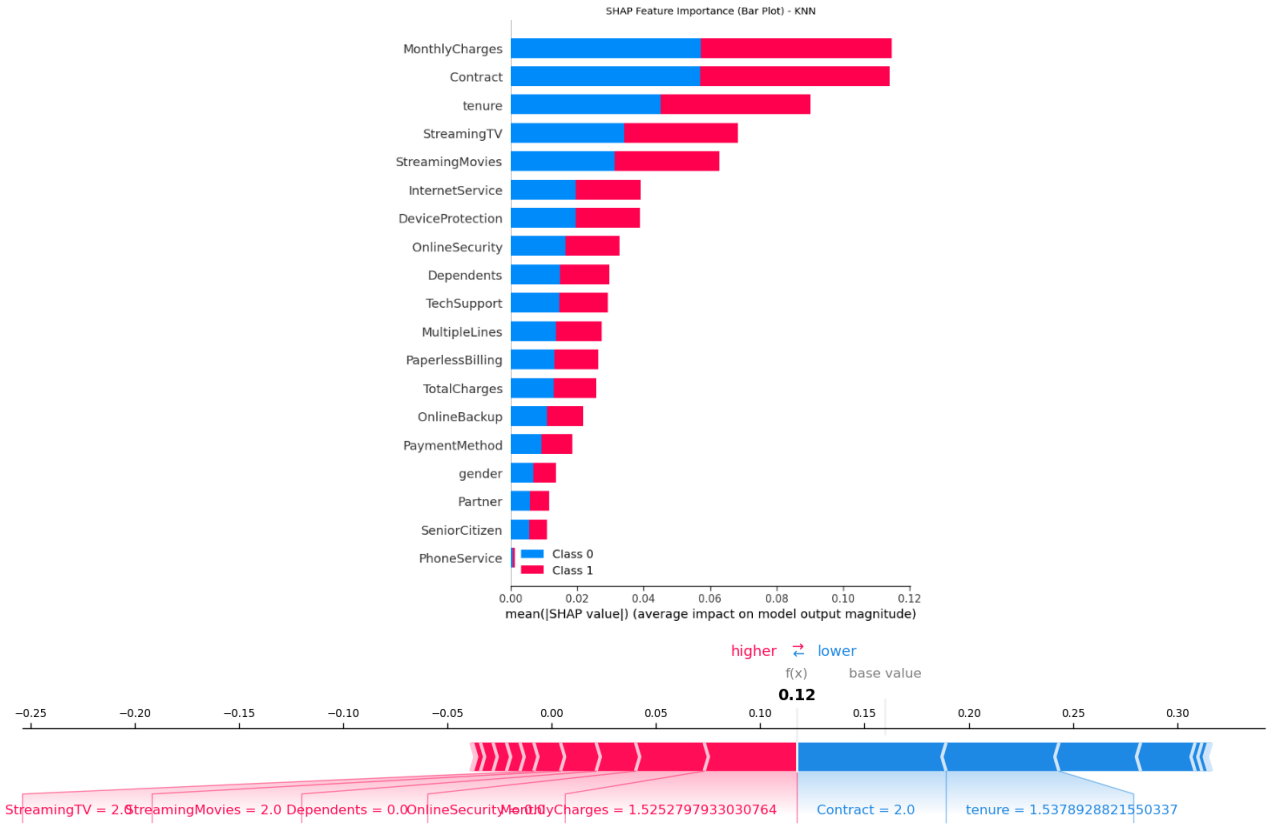

Fig.12 SHAP analysis of KNN

The SHAP interpretability shows that the most influencing factor of churn are the features like MonthlyCharges, Contract, and Tenure, while other features like StreamingTV and StreamingMovies have slight influence on the predictions. These information helps the companies to manage the contract to reduce the charges, provide better services and developing personalised plan for making the customer stay.

6.4 Experiment 4: Multi-stacked Model

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

           0       0.84      0.92      0.88      1549
           1       0.70      0.53      0.60       561

    accuracy                           0.81      2110
   macro avg       0.77      0.72      0.74      2110
weighted avg       0.80      0.81      0.80      2110
```

Fig.13 Classification Report of Multi-Stacked model

The ouputs of the base models such as Random Forest, XGBoost and KNN are provided as inputs for the multi-stacked model or meta model with Logistic Regression. In Fig.13, the evaluation of multi-stacked models displays robust predictive performance by achieving an accuracy of 81% comparatively more than all the base models, with large number of No Churn class showing a precision of 84%, recall of 92% and F1-score of 88%. For Churn class it showed a precision of 70%, recall of 53% and F1-score of 60%, displaying better performance than other models. Also, the ROC score showed high score of 0.85.

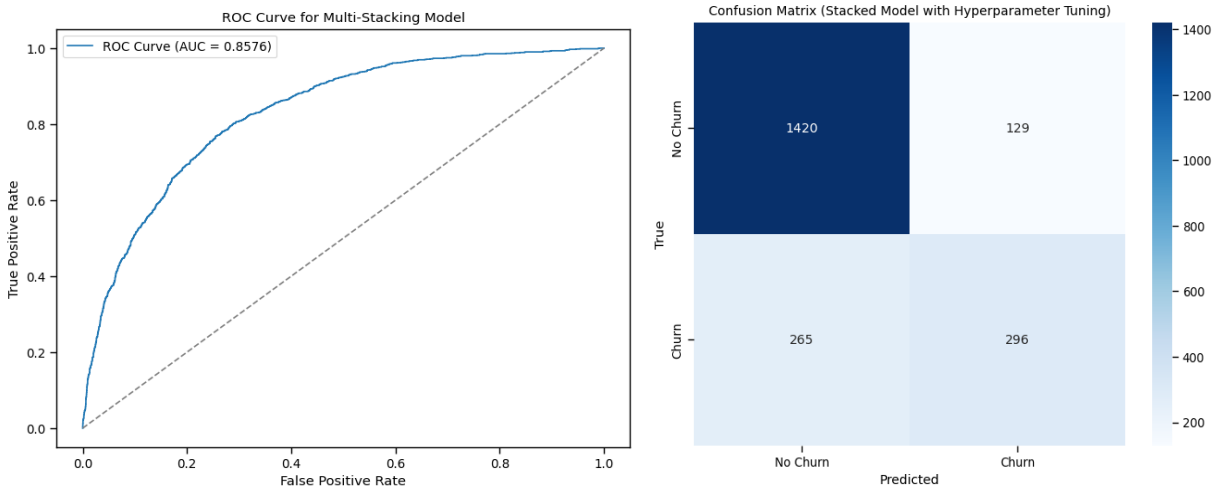

Fig.14 ROC-AUCcurve and Confusion Matrix for Multi-stacled model

The confusion matrix revealed 1420 True negatives that is No Churn labelled correctly and around 296 Churn customers were labelled correctly (True positives). Also 129 customers were misidentified as Churn (false positives) and 254 customers were identified as No Churn (False negatives). Overall, the multi-stacked ensemble model displayed better performance when compared to all base models individually.

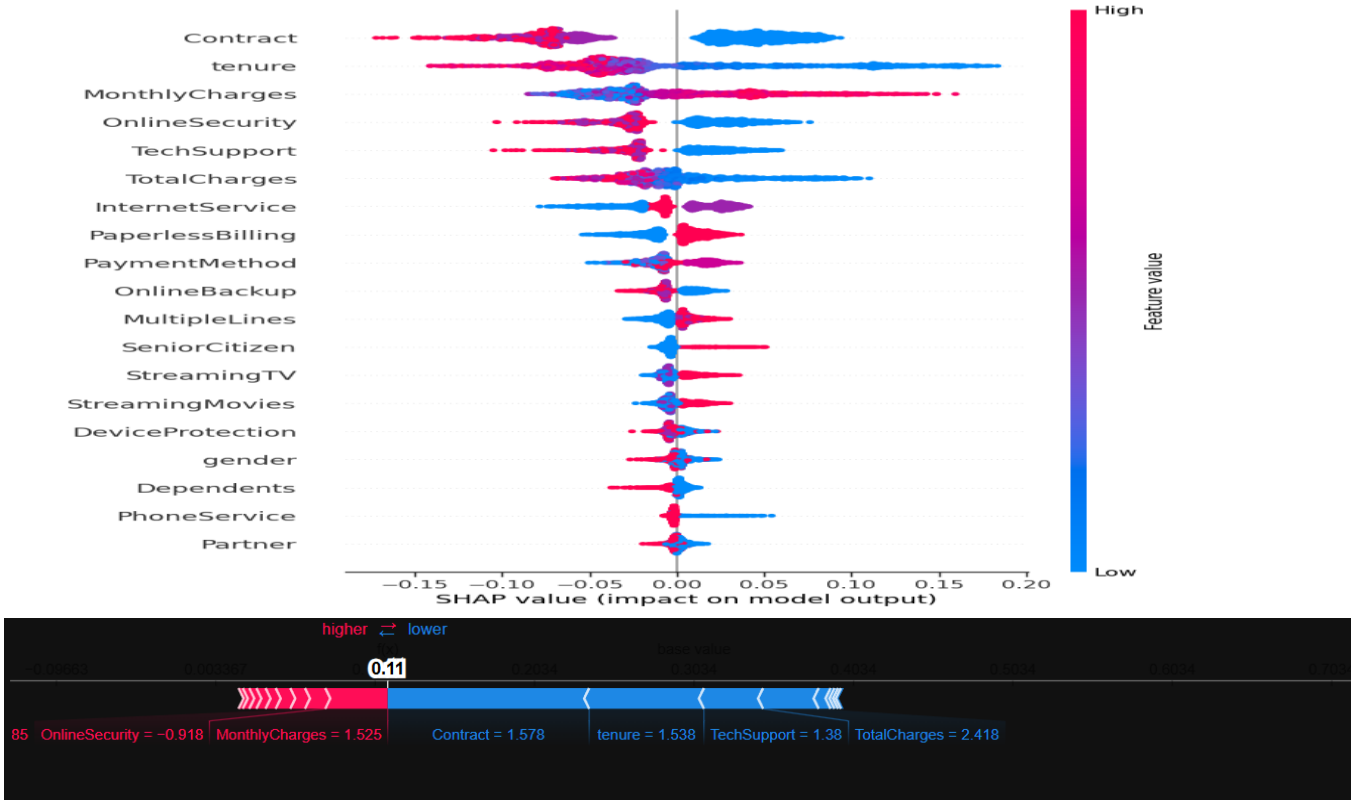

Fig.15 SHAP analysis of Multi-Stacked Model

Like SHAP analysis done on the base models, the analysis on multi-stacked model shows that longer contract and tenure reduces churn risk while high charges increase the risk of churn by identifying the significant features like MonthlyCharges, Tenure and Contract which are influencing the churn rate. Some feature like Online Security and TechSupport also have their influence. The individual churning factors are provided Local explanations which help the companies to design targeted strategies by improving services, offering discounts and promoting long-term contracts this will help to improve customer retention and satisfaction.

6.5 Results and Discussion

Table.2 Comparison of Model's Performance

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Macro Avg F1-Score
Random Forest	0.76	0.55	0.70	0.61	0.72
Tuned XGBoost	0.75	0.53	0.66	0.58	0.70
Tuned KNN	0.79	0.60	0.58	0.59	0.72
Stacked Model with Hyperparameter Tuning	0.81	0.70	0.53	0.6	0.74

The study proposes a multi-stack model by using the results of various base models as input for the meta model to improve the efficiency of customer churn prediction and enhancing the interpretability by applying Explainable AI (XAI) technique called SHapley Additive exPlanations (SHAP). The experiments on various models provided significant insights on churn prediction performance. Random Forest one of the base models displayed an accuracy of 76%, with a recall of 70%, however the model showed a low precision of 55% and a F1-score of 61%. There was step down in the performance of the second base model, XGBoost as it achieved an accuracy of 75% with a precision of 53%, recall of 66%, and F1-score of 58%. The third base model performed better than the other two base models with an accuracy of 79% with high precision of 60%, nonetheless there was a dip in recall value of 59%. The model stacked with multiple base models with logistic regression performed best with an 81% with high F1-score of 74%, precision value of 70%. Still the model paid the price for identifying lesser false positives with recall value of 53%. However, the implementation of multi-stacked approach by integrating strengths various ensemble models has enhanced performance in customer churn predictions.

The SHAP analysis across all the models revealed that features like Contract, MonthlyCharges, and Tenure were the significant factors influencing the churn. The analysis showed the importance of financial and contractual factors as the customers with long term contract and long tenure were less likely even when the prices increased to the customers with short term contract and tenure. Additionally, customers who did not use services like OnlineSecurity and TechSupport more likely to churn. The influence of the monthly charges and value-added services on customer churn was provided by the personalised insights of SHAP force plot.

Although, the models seem to be reliable, it leaves room for improvement. Even though its common to use SMOTE for handling the class imbalance, it may introduce noise. So, it may be better to apply SMOTE-ENN or Tomek Links. Also, the application of RFE for feature selection was useful, but the implementation of RFE integrated with LASSO could improve the results. The implementation of Logistic Regression as stacked ensemble model could not capture the complex relations which may be possible by applying Gradient Boosting or neural network models. The scalability of the model may be affected by implementing only one dataset and the model need to be tested on real-time telecom data. The SHAP analysis provided proper explanation global and local feature importance, but its computational cost may cause problem of scalability on large dataset, so the integration of SHAP and LIME could help to provide efficient interpretability at a respectable cost.

7 Conclusion and Future Work

The study successfully met its objective by addressing challenge of customer churn predictions in the telecommunication industry by improving the accuracy and interpretability. Th research focused on employing advance machine learning like Random Forest, XGBoost, KNN as base model and Logistic Regression as meta model for the development of multi-stacked ensemble model. To handle the class imbalance SMOTE technique was used and feature selection was performed by using Recursive Feature Elimination method. The implementation SHAP enhanced the interpretability, and it also helped identify the key metrics driving the churn. The multi-stacked framework outperformed all the base models by achieving an accuracy of 81% with Macro Average of F1-score of 74% demonstrating robust performance. The application of SHAP confirmed the importance of features like Tenure, MonthlyCharges and Contract as the factors influencing the churn. These insights can help the companies to trust the results of the model and management to design customer retention strategies.

In the future work, the oversampling noise could be decreased by the application advance resampling techniques like SMOTE-ENN and implementation of combination of RFE and LASSO for better feature selection. Expand the dataset by including real-world and complex datsets that would help to increase the scalability and reliability. Implementation of neural networks like RNN and LSTM and building a multi-stacked model using LightGBM, CatBoost or Gradient Boosting, and also performing advance hyperparameter search, could help in enhancing the results. Integration Explainable AI models like LIME and SHAP may result in enhancing the results. The proposed framework shows positive growth towards the commercial use and adaptability of the real-world telecommunication data that would help companies lower the churn and design better customer retention strategies.

References

Aggarwal, P. and Vijayakumar, V., 2024, May. Customer Churn Prediction in the Telecom Sector. In *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)* (pp. 1-6). IEEE.

Ahmad, A.K., Jafar, A. and Aljoumaa, K., 2019. Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), pp.1-24.

Barsotti, A., Gianini, G., Mio, C., Lin, J., Babbar, H., Singh, A., Taher, F. and Damiani, E., 2024. A Decade of Churn Prediction Techniques in the TelCo Domain: A Survey. *SN Computer Science*, *5*(4), p.404.

Bharambe, Y., Deshmukh, P., Karanjawane, P., Chaudhari, D. and Ranjan, N.M., 2023, January. Churn prediction in telecommunication industry. In *2023 International Conference for Advancement in Technology (ICONAT)* (pp. 1-5). IEEE.

Chang, V., Hall, K., Xu, Q.A., Amao, F.O., Ganatra, M.A. and Benson, V., 2024. Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, *17*(6), p.231.

Faraji Googerdchi, K., Asadi, S. and Jafari, S.M., 2024. Customer churn modeling in telecommunication using a novel multi-objective evolutionary clustering-based ensemble learning. *Plos one*, *19*(6), p.e0303881.

He, C. and Ding, C.H., 2024. A novel classification algorithm for customer churn prediction based on hybrid Ensemble-Fusion model. *Scientific Reports*, *14*(1), p.20179.

Kavitha, C., Tripathi, G., Sridivya, R., Yamuna, V., Supriya, N. and Lakshmanarao, A., 2024, April. An Efficient Churn Prediction model using ML supervised and semi supervised Learning Techniques. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1-6). IEEE.

Kirgiz, O.B., Kiygi-Calli, M., Cagliyor, S. and El Oraiby, M., 2024. Assessing the effectiveness of OTT services, branded apps, and gamified loyalty giveaways on mobile customer churn in the telecom industry: A machine-learning approach. *Telecommunications Policy*, *48*(8), p.102816.

Ouf, S., Mahmoud, K.T. and Abdel-Fattah, M.A., 2024. A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry. *Journal of Big Data*, *11*(1), p.70.

Poudel, S.S., Pokharel, S. and Timilsina, M., 2024. Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications*, *17*, p.100567.

Shaikhsurab, M. a. M. P., 2024. Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach. *arXiv preprint*, Volume arXiv:2408.16284.

Sharma, A., Tyagi, M. and Kumar, S., 2023, June. Telecom Churn Prediction Using Python. In *International Conference on Recent Trends in Computing* (pp. 423-434). Singapore: Springer Nature Singapore.

Sikri, A., Jameel, R., Idrees, S.M. and Kaur, H., 2024. Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, *14*(1), p.13097.

Usman-Hamza, F.E., Balogun, A.O., Amosa, R.T., Capretz, L.F., Mojeed, H.A., Salihu, S.A., Akintola, A.G. and Mabayoje, M.A., 2024. Sampling-based novel heterogeneous multi-layer stacking ensemble method for telecom customer churn prediction. *Scientific African*, *24*, p.e02223.

Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P. and Gawande, S.H., 2024. Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, *14*, p.100342.

Wang, C., Rao, C., Hu, F., Xiao, X. and Goh, M., 2024. Risk assessment of customer churn in telco using FCLCNN-LSTM model. *Expert Systems with Applications*, *248*, p.123352.