Predictive Modeling and Feature Analysis for Customer Churn in Consumer-Facing Industries

Md. Fahim Shakil Chowdhury
Department of CSE
20210104128
Dhaka, Bangladesh
fahim.cse.20210104128@aust.edu

Sanjana Akhter
Department of CSE
20210104150
Dhaka, Bangladesh
sanjana.cse.20210104150@aust.edu

Md. Mostafizur Rahman
Department of CSE
20210104151
Dhaka, Bangladesh
mostafizur.cse.20210104151@aust.edu

***Abstract*—Customer churn prediction is a critical challenge for businesses, especially in the telecommunications industry where high competition makes customer retention essential. This study uses the Telco Customer Churn dataset containing 7,043 records with 21 attributes to identify factors influencing churn and to build predictive models. Data preprocessing involved handling missing values, encoding categorical features, and normalizing numeric variables. Feature importance was determined through Random Forest, XGBoost, and LightGBM, leading to the selection of top feature subsets. Two pipelines were developed: one using class weighting and another with SMOTE-NC oversampling to address class imbalance. Logistic Regression, Random Forest, XGBoost, and LightGBM models were trained with hyperparameter tuning, along with hard and soft ensemble strategies. Performance was evaluated using accuracy, precision, recall, specificity, and AUC-ROC. The results demonstrate that ensemble methods, particularly with SMOTE-NC, achieved the most stable and accurate churn predictions, and that reduced feature sets (Top 15 and Top 11) often outperformed the full dataset. This research highlights the effectiveness of ensemble learning, feature selection, and oversampling techniques in reducing customer attrition and improving decision-making for retention strategies.**

***Index Terms*—Customer churn, Telco dataset, Logistic Regression, Random Forest, XGBoost, LightGBM, Ensemble models, SMOTE-NC, Predictive analytics, Feature selection**

I. Introduction

In today's highly competitive market, customer retention has become one of the most significant challenges for businesses, especially in consumer-facing industries such as telecommunications, retail, and services. Customer churn, which occurs when customers discontinue or cancel their subscription to a service, has significant financial implications for companies. The loss of customers not only leads to reduced revenue but also increases the cost of acquiring new customers to replace those lost.

For businesses, understanding the reasons behind customer churn is critical, as it allows them to design targeted strategies to retain customers before they decide to leave. Predicting which customers are at risk of churning enables businesses to take proactive measures and offer personalized interventions, such as promotions, loyalty programs, or customer support enhancements, aimed at improving customer satisfaction and retention.

By identifying early warning signs and the key factors driving churn, companies can reduce customer attrition rates, increase Customer Lifetime Value (CLV), and ultimately enhance overall profitability. This predictive modeling approach not only provides businesses with the tools to forecast customer behavior but also empowers them to act before it's too late, making it a valuable tool in maintaining long-term customer relationships and sustained business success.

II. Literature Review

[1]**Telco Customer Churn Prediction** This study by Ying Wei explores customer churn prediction in the telecommunication industry using the Telco Customer Churn dataset (7043 rows, 21 features). It compares Classification Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression models. Logistic Regression achieved the highest performance with an accuracy of 81.7% and an AUC score of 0.852, followed by Random Forest (81.17% accuracy, 0.837 AUC). The dataset, with 5174 non-churning and 1869 churning customers, required preprocessing (handling missing values, duplicates, removing 'customer ID', and transforming categorical data). PCA and stepwise regression were applied for feature selection. Logistic Regression excelled due to its efficiency in predicting churn, aiding pricing strategies and service quality improvements.

[2]**Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Model** Victor Chang and colleagues investigated churn prediction using the "Telecom CUSTOMER Churn" dataset (7043 rows, 38 features). Models evaluated include Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, and K-Nearest Neighbor (KNN). Random Forest outperformed others with 86.94% accuracy and 0.95 AUC, while Logistic Regression scored 75.53% accuracy and 0.84 AUC. The imbalanced

dataset (twice as many churning samples) underwent preprocessing, including handling missing values and transforming categorical data. Explainable AI tools (LIME, SHAP) and CRISP-DM methodology were used for interpretability and analysis. Random Forest provided actionable insights for customer retention strategies.

[3]**Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform** Ahmad et al. (2019) developed a churn prediction model for SyriaTel using a 70-terabyte dataset on Hortonworks Data Platform with Spark. Models tested include Decision Tree (83% AUC), Random Forest (87.76% AUC), GBM (90.89% AUC), and XGBoost (93.3% AUC). XGBoost performed best, leveraging statistical (tenure, balance) and SNA features (MTN Cosine similarity, inactivity). Without rebalancing the imbalanced dataset (5% churners), XGBoost achieved 89% AUC on 7.5 million customers, reducing churn by 1.5% through proactive retention.

[4]**Causal Analysis of Customer Churn Using Deep Learning** This study presents a framework for churn prediction and causal analysis using high-dimensional financial data from superannuation funds. A Deep Feedforward Neural Network (DFF NN), combined with SMOTE and ensemble methods, matched top classifiers, with XGBoost achieving the highest AUC (80%). Bayesian networks identified key churn factors (recent contributions, account growth, balance amounts). The approach enabled a potential 3% churn reduction among active customers with accounts over a year old, supporting targeted retention strategies.

[5]**Customer Churn Prediction Model using Explainable Machine Learning** This study develops a churn prediction model using XGBoost, achieving 88% accuracy, with precision (55%), recall (22%), and F1-score (32%). SHAP values enhanced model interpretability by explaining feature contributions, increasing stakeholder trust and transparency.

III. Data Analysis

The dataset employed in this study is the Telco Customer Churn dataset, consisting of 7,043 customer records with 21 attributes. Each record contains demographic information (gender, age indicator, partner and dependents), service usage details (phone, internet, security add-ons, streaming), account information (tenure, contract type, billing method, payment method), and the target variable Churn indicating whether the customer discontinued the service.

Figure 1 shows the feature and the correlation between them.

Figure 2 shows Correlation of features with churn

Correlation analysis revealed that churn is negatively correlated with tenure and contract length, while positively correlated with electronic check payment method and paperless billing. Moreover, service add-ons such as OnlineSecurity and TechSupport showed a protective effect, indicating that customers with these features are less likely to churn.

The analysis reveals that different features influence churn rates in distinct ways. By examining the distributions, it becomes evident which attributes play a more critical role in driving customer attrition. Figure 3 shows the Churn rate by contract

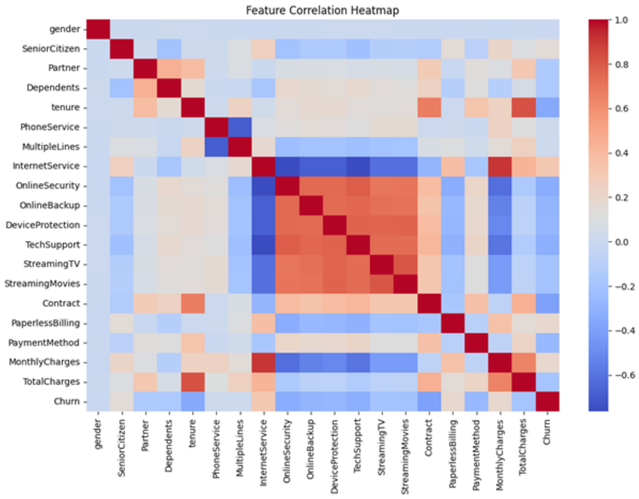

Fig. 1: Feature and the corelation between them

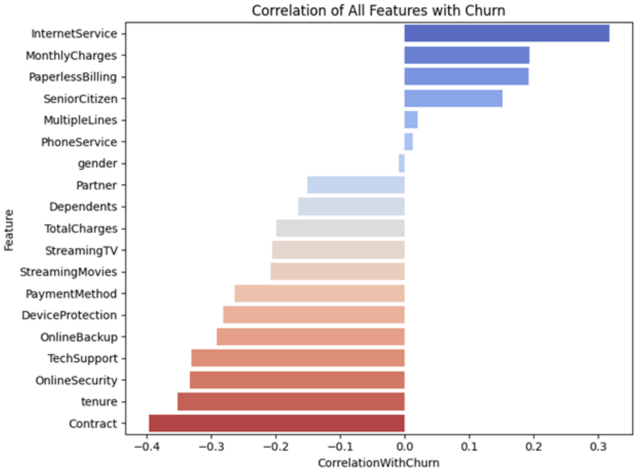

Fig. 2: Corelation of features with churn

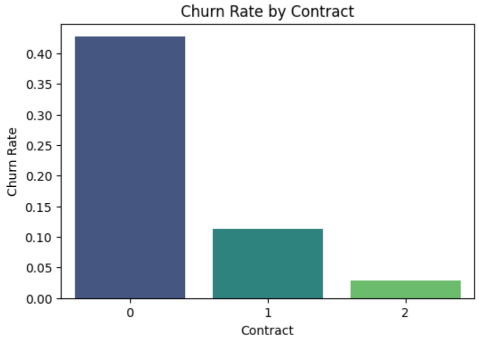

Fig. 3: Churn rate by contract (0-Month, 1-Year and 2-Year)

For example, customers on month-to-month contracts exhibit significantly higher churn compared to those on longer-term contracts, while the type of internet service also shows a strong relationship with churn likelihood. These patterns highlight that churn is not uniform across all features but is shaped by specific customer and service characteristics, making them key factors to focus on for effective retention strategies.

Initially, all column names and string entries were stripped of trailing spaces to ensure consistency. The unique customer identifier was removed as it does not contribute to prediction. Missing values were examined; the attribute TotalCharges contained null entries which were replaced with the median value to preserve distribution.

Continuous variables (tenure, MonthlyCharges, TotalCharges) were normalized using z-score standardization, ensuring that all features contributed comparably during model training. The SeniorCitizen attribute was standardized to integer binary format (0/1). Binary categorical variables (e.g., Partner, Dependents, PhoneService, PaperlessBilling) were mapped into 0/1 encoding. Multi-level service features (e.g., MultipleLines, OnlineSecurity, TechSupport) were encoded into label categories (0 = No, 1 = Yes, 2 = No service). Gender was mapped to 0/1. The multi-class attributes InternetService, Contract, and PaymentMethod were label and ordinally encoded. The target variable Churn was converted into binary form (0 = No, 1 = Yes).

Figure 4 shows the distribution of the class of the dataset.

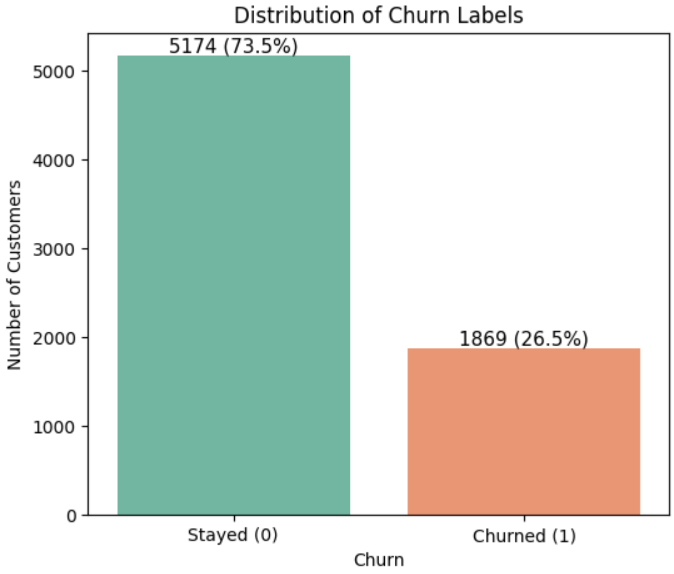

Fig. 4: Dataset Class Distribution

The churn distribution was imbalanced, with the majority of customers labeled as “ No Churn.” To address this, SMOTE-NC (Synthetic Minority Oversampling Technique for Nominal and Continuous features) was applied. This technique generates synthetic samples of minority churn cases while respecting categorical feature distributions, thereby mitigating class imbalance and improving model generalizability.

Thus, after cleaning, transformation, correlation inspection, and class rebalancing using SMOTE-NC, the dataset was standardized and prepared for feature selection and subsequent model development.

IV. Methodology

After data preprocessing , the implementation was done using several machine learning algorithms.

A. Feature Engineering and Selection

To improve model performance and reduce complexity, a robust feature selection process was implemented with the aim of identifying the most influential features for predicting churn. Three powerful tree-based models—Random Forest, XGBoost, and LightGBM—were trained on the dataset, and each produced an importance score for every feature. These scores were then collected, normalized to a common scale (0–1) to ensure fair comparison, and averaged to compute a final MeanScore for each feature. The features were subsequently ranked according to their MeanScore, providing an ensemble-based ranking that reduces bias toward any single model’s preferences. Based on this ranking, two subsets of features were defined for experimentation: the Top 15 and Top 11 most important features. Figure 5 shows Combined Feature importance

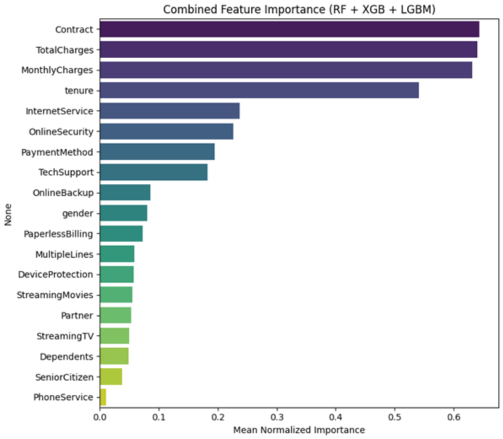

Fig. 5: Combined Feature importance

B. Model Development and Evaluation

Two parallel pipelines were established to compare different strategies for handling the dataset’s inherent class imbalance (fewer ’Churn’ instances than ’No Churn’). For both pipelines, the data was split into an 80% training set and a 20% testing set using stratified sampling to maintain the original churn distribution in both splits. Figure 6 shows Implementation pipeline

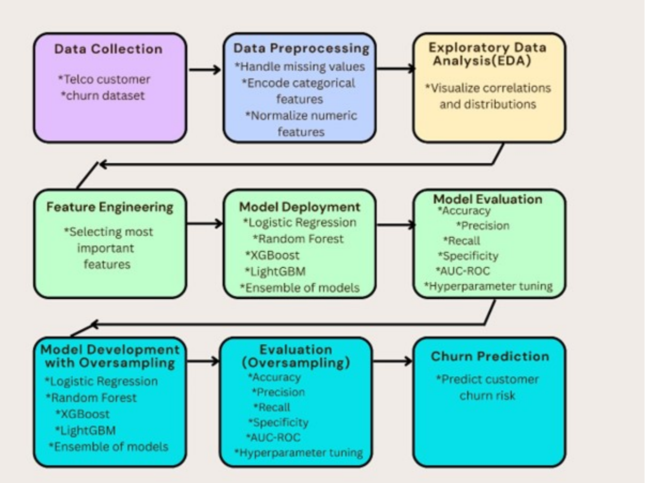

Fig. 6: Implementation pipeline

Pipeline A: Imbalance Handling with Model Weighting:

- **Imbalance Handling**
  - Each model was configured to penalize misclassification of the minority class more heavily.
  - For boosting algorithms (XGBoost, LightGBM), the `scale_pos_weight` parameter was applied.
  - For Random Forest and Logistic Regression, class weights were balanced accordingly.
- **Models Trained**
  1) Logistic Regression (tuned hyperparameters)
  2) Random Forest (tuned hyperparameters)
  3) XGBoost (tuned hyperparameters)
  4) LightGBM (tuned hyperparameters)
- **Experiments**
  - Each model was trained and evaluated on three distinct feature sets:
    – All Features (full dataset)
    – Top 15 Features (based on model-aggregated importance ranking)
    – Top 11 Features (further refined subset)
- **Ensemble Models**
  - After evaluating individual models, two ensemble strategies were applied:
    – **Hard Voting Ensemble**: majority rule on predicted class labels.
    – **Soft Voting Ensemble**: average of predicted probabilities.
  - Ensembles combined Logistic Regression, Random Forest, XGBoost, and LightGBM to improve predictive accuracy and stability.

Pipeline B: Imbalance Handling with Oversampling (SMOTE-NC):

- **Imbalance Handling**
  - The training data was balanced directly using the Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC).
  - SMOTE-NC was chosen because it can generate synthetic samples for datasets containing both categorical and continuous variables.
  - The technique was applied only to the training split inside each cross-validation fold and before the final model training, ensuring no data leakage into validation or test sets.
- **Models Trained**
  1) Logistic Regression (tuned hyperparameters)
  2) Random Forest (tuned hyperparameters)
  3) XGBoost (tuned hyperparameters)
  4) LightGBM (tuned hyperparameters)
- **Experiments**
  - Each model was trained and evaluated on three distinct feature sets:
    – All Features (full dataset)
    – Top 15 Features (based on model-aggregated importance ranking)
    – Top 11 Features (further refined subset)
- **Ensemble Models**
  - After evaluating individual models, two ensemble strategies were constructed:
    – **Hard Voting Ensemble**: majority rule on predicted class labels.
    – **Soft Voting Ensemble**: average of predicted probabilities.
  - These ensembles combined Logistic Regression, Random Forest, XGBoost, and LightGBM trained on the SMOTE-NC balanced data to further improve accuracy and stability.

C. Evaluation Metrics

To comprehensively assess model performance on an imbalanced dataset, multiple evaluation metrics were employed :

- **Accuracy :** The ratio of correctly predicted observations to the total number of observations. Although intuitive, accuracy alone may be misleading in imbalanced settings.
- **Precision :** The proportion of correctly predicted churn cases among all customers predicted as churners, i.e.This metric reflects how reliable the model is when it predicts churn.
- **Recall (Sensitivity) :** The proportion of actual churners correctly identified by the model, i.e.Recall is critical in churn prediction, as failing to detect churners may lead to revenue loss.
- **F1-Score :** The harmonic mean of Precision and Recall, defined as It provides a balanced single measure, particularly useful under class imbalance.
- **Specificity :** The proportion of non-churners correctly identified, i.e.Specificity complements Recall by ensuring that customers who are unlikely to churn are not misclassified.
- **AUC-ROC :** The Area Under the Receiver Operating Characteristic Curve. This metric evaluates the model's

discriminative ability across varying thresholds, where higher values indicate better separation between churners and non-churners.

- **Average Precision (PR-AP) :** The area under the Precision–Recall curve, particularly useful in imbalanced scenarios since it emphasizes performance on the minority (churn) class.

V. Result Analysis

A. *Baseline Performance (Before Oversampling)*

Table 1 summarizes the performance of all four models (Logistic Regression, Random Forest, XGBoost, LightGBM) across three feature sets (All Features, Top 15, Top 11), along with the Hard Voting and Soft Voting ensembles. The reported metrics include Accuracy, Precision, Recall, F1-score, and Specificity. **Table 2** summarizes the performance of all four models (Logistic Regression, Random Forest, XGBoost, LightGBM) across three feature sets (All Features, Top 15, Top 11), along with the Hard Voting and Soft Voting ensembles. The reported metrics include Accuracy, Precision, Recall, F1-score, and Specificity.

TABLE I: Performance of models before oversampling

Model	Acc.	Prec.	Rec.	F1	Spec.
LR All	0.7381	0.7047	0.7551	0.7087	0.7188
LR Top 15	0.7331	0.7023	0.7525	0.7047	0.7101
LR Top 11	0.7253	0.6966	0.7473	0.6974	0.7005
RF All	0.7502	0.7145	0.7659	0.7205	0.7320
RF Top 15	0.7495	0.7140	0.7654	0.7199	0.7310
RF Top 11	0.7559	0.7164	0.7655	0.7243	0.7450
XGB All	0.7495	0.7145	0.7663	0.7202	0.7300
XGB Top 15	0.7459	0.7111	0.7622	0.7163	0.7280
XGB Top 11	0.7431	0.7100	0.7619	0.7143	0.7220
LG All	0.7573	0.7192	0.7699	0.7267	0.7430
LG Top 15	0.7523	0.7147	0.7648	0.7216	0.7380
LG Top 11	0.7559	0.7164	0.7655	0.7243	0.7450
HVE All	0.7488	0.7149	0.7675	0.7201	0.7270
HVE Top 15	0.7438	0.7115	0.7641	0.7155	0.7210
HVE Top 11	0.7480	0.7140	0.7662	0.7191	0.7210
SVE All	0.7495	0.7150	0.7671	0.7204	0.7290
SVE Top 15	0.7459	0.7125	0.7647	0.7172	0.7246
SVE Top 11	0.7495	0.7159	0.7688	0.7210	0.7275

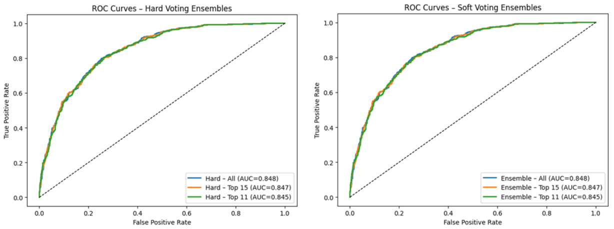

Fig. 7: ROC curves of the Hard Voting and Soft Voting ensembles (baseline, before SMOTE).

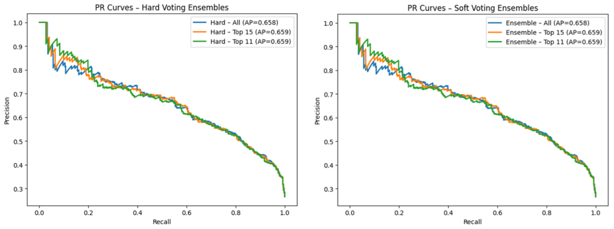

Fig. 8: Precision–Recall curves of the Hard Voting and Soft Voting ensembles (baseline).

B. *Performance After Oversampling (SMOTE-NC)*

TABLE II: Performance of models after SMOTE-NC oversampling

Model	Acc.	Prec.	Rec.	F1	Spec.
LR All Features	0.7999	0.7437	0.7297	0.7360	0.8050
LR Top 15	0.7935	0.7350	0.7220	0.7278	0.8000
LR Top 11	0.7913	0.7321	0.7205	0.7258	0.7950
RF All Features	0.8034	0.7506	0.7210	0.7329	0.8100
RF Top 15	0.8006	0.7464	0.7174	0.7290	0.8050
RF Top 11	0.7999	0.7462	0.7126	0.7256	0.8000
XGB All Features	0.8020	0.7485	0.7192	0.7309	0.8070
XGB Top 15	0.8055	0.7540	0.7225	0.7350	0.8100
XGB Top 11	0.7984	0.7431	0.7159	0.7269	0.8020
LG All Features	0.7999	0.7457	0.7152	0.7272	0.8050
LG Top 15	0.8020	0.7491	0.7167	0.7293	0.8070
LG Top 11	0.7984	0.7428	0.7177	0.7280	0.8030
HVE All Features	0.7977	0.7412	0.7223	0.7304	0.8830
HVE Top 15	0.8013	0.7468	0.7221	0.7324	0.8910
HVE Top 11	0.8041	0.7538	0.7138	0.7287	0.9060
SVE All Features	0.7956	0.7378	0.7285	0.7328	0.8710
SVE Top 15	0.8034	0.7497	0.7253	0.7355	0.8920
SVE Top 11	0.8077	0.7603	0.7154	0.7316	0.9120

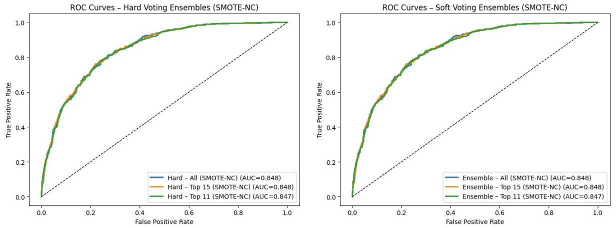

Fig. 9: ROC curves of the Hard Voting and Soft Voting ensembles (baseline, after SMOTE).

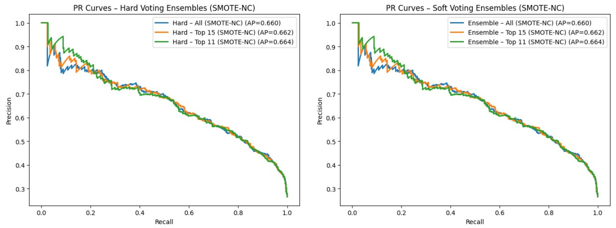

Fig. 10: Precision–Recall curves of the Hard Voting and Soft Voting ensembles (baseline,after SMOTE).

C. *Overall Comparative Analysis*

Before oversampling, the models showed reasonable performance with 73–76% accuracy and F1 scores around 0.69–0.72, generally favoring recall (0.75–0.77) over precision (0.70), which means they were good at catching churners but at the cost of many false positives. Specificity values were modest, mostly 0.72–0.75, confirming this tendency to misclassify non-churners. After applying SMOTE-NC oversampling, the results became more balanced and reliable: accuracy increased to 79–81%, F1 improved to 0.73–0.74, and specificity rose significantly to 0.80–0.91, showing fewer false alarms. Notably, ensemble methods, especially the Soft Voting Ensemble with Top 15 or Top 11 features, consistently outperformed individual models by combining high accuracy, stable recall–precision balance, and excellent specificity, making them the most effective approach for churn prediction in this study.

VI. Conclusion and Future Work

This study demonstrates that machine learning provides a powerful and reliable framework for predicting customer churn in the telecommunications sector. Logistic Regression, Random Forest, XGBoost, and LightGBM—individually and as ensembles—delivered strong predictive performance, with ensemble strategies and SMOTE-NC oversampling producing the most stable and accurate results. Importantly, experiments revealed that reduced feature sets (Top 15 and Top 11) often performed better than the full dataset, underscoring the value of feature selection. This not only improved efficiency but also highlighted the most influential drivers of churn, giving businesses actionable insights into which customer attributes deserve closer attention. The analysis also confirmed that different churn-related features, such as contract type, payment method, and support services, influence attrition in different ways, reinforcing the need for tailored retention strategies. Overall, the findings establish that a combination of rigorous preprocessing, feature selection, and ensemble learning forms an effective blueprint for churn prediction and customer retention.

For Future work :

- **Dataset enrichment :** Expanding to larger or more diverse datasets from other industries to test the model's generalizability.
- **Dynamic churn modeling :** Exploring time-series approaches that capture how churn risk evolves over a customer's lifecycle.
- **Advanced ensemble methods :** Investigating stacking or hybrid deep learning models to further improve performance.

References

[1] Y. Wei, ”Telco Customer Churn Prediction,” in Highlights in Science, Engineering and Technology SDPIT 2024, vol. 92, 2024 Available: ResearchGate.

[2] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, “Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models,” Algorithms, vol. 17, Art. ID 231, May 2024, doi: 10.3390/a17060231. Available: MDPI.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” J. Big Data, vol. 6, no. 28, pp. 1-24, 2019, doi: 10.1186/s40537-019-0191-6 . Available: arXiv.

[4] D. H. Rudd, H. Huo and G. Xu, ”Causal Analysis of Customer Churn Using Deep Learning,” arXiv preprint, arXiv:2304.10604, 2023. Available: arXiv.

[5] 5. J. Maan and H. Maan, “Customer Churn Prediction Model using Explainable Machine learning,” International Journal of Computer Science Trends and Technology (IJCST), vol. 11, no. 1, pp. 33-38, Jan.-Feb. 2023 . Available: arXiv.