Predicting Churn of Telecommunications Customers

Margaret G. Brier
Student, Master's of Computer Science
College of Computing and Informatics
Drexel University
mgb87@drexel.edu

Yash Gupta
Student, Master's of Computer Science
College of Computing and Informatics
Drexel University
yg444@drexel.edu

Shubham Jadhav
Student, Master's of Computer Science
College of Computing and Informatics
Drexel University
sj3237@drexel.edu

***Abstract*—This paper explores the utilization of logistic regression and decision trees, along with an ensemble approach using random forests, to predict customer churn in the telecommunications sector. With the industry facing significant challenges due to high churn rates, effective prediction models are essential for identifying at-risk customers and implementing strategic retention efforts. We conducted an exploratory data analysis to identify relevant features and patterns, followed by the application of logistic regression to estimate the probability of churn based on customer data. Additionally, decision trees were developed to further dissect the churn dynamics, incorporating a novel feature engineering and selection process to enhance model accuracy. The study also introduces an ensemble method through the deployment of random forests, aiming to capitalize on the strengths of multiple decision trees to improve prediction outcomes. Additionally we did comparisons between all the models. Our findings reveal that each model offers distinct advantages in understanding and predicting churn, with logistic regression providing clear probabilistic outputs and decision trees offering in-depth insights into customer behavior. The random forest method demonstrated superior performance, combining the predictive power of numerous trees to achieve greater accuracy and robustness in churn prediction. This research underscores the potential of machine learning techniques in aiding telecommunications companies to better understand customer churn, thereby enabling more effective retention strategies and contributing to sustained business growth.**

***Index Terms*—logistic regression, decision tree, random forest, ensemble, feature reduction, machine learning**

I. Introduction

In the competitive landscape of the telecommunications industry, customer churn—where customers cease their relationships with a company—presents a significant challenge, impacting revenue and long-term growth. Predicting which customers are likely to churn enables companies to deploy targeted interventions, aiming to retain valuable customers and enhance satisfaction. This paper presents a comprehensive study on the application of logistic regression and decision trees, along with an ensemble method using random forests, to predict customer churn within a telecommunications context.

We begin by exploring the background and related work in the field, highlighting the importance of churn prediction and reviewing previous approaches. Our methodology is rooted in an exploratory data analysis to understand the characteristics and patterns within the data, followed by the application of logistic regression to model the probability of churn based on various customer features. We further explore decision trees for churn prediction, employing a novel approach to feature engineering and selection to enhance model performance.

Our results demonstrate the efficacy of both logistic regression and decision trees in predicting churn, with each model offering unique insights into the factors influencing customer decisions. Through a detailed analysis, we identify key predictors of churn and assess the models' performance using various metrics. The addition of a random forest ensemble method further refines our predictions, leveraging the strengths of multiple decision trees to improve accuracy and robustness.

The contributions of this work are twofold: First, we provide a detailed comparison of logistic regression and decision trees in the context of churn prediction, including an examination of feature importance and model interpretability. Second, we demonstrate the value of ensemble methods, specifically random forests, in enhancing prediction accuracy and providing deeper insights into complex customer behavior patterns.

Lastly, we innovatively combine the predictive strengths of logistic regression and decision trees into a powerful ensemble model, utilizing random forests. This ensemble approach not only amplifies the predictive accuracy by aggregating diverse decision pathways but also mitigates the individual limitations of each model, leading to a more robust prediction of customer churn. The integration of these methodologies embodies a strategic fusion of linear and non-linear prediction mechanisms, offering a comprehensive tool for telecommunications companies to proactively manage and reduce customer churn. Through this ensemble model, we significantly enhance the ability to identify at-risk customers, thereby enabling more targeted and effective retention strategies.

Finally our study offers valuable perspectives for telecommunications companies seeking to reduce churn and foster customer loyalty. By harnessing the power of logistic regression, decision trees, and random forests, businesses can more effectively predict and mitigate the risk of customer departure, securing a competitive edge in a dynamic market.

II. Background and Related Work

The telecommunications industry stands at the forefront of digital transformation, with the advent of new technologies continually reshaping consumer behaviors and market dynamics. Amidst this rapid evolution, customer churn remains a persistent challenge, exerting significant pressure on telecom companies to devise effective retention strategies. To address this issue, researchers have increasingly turned to machine learning methodologies, leveraging the wealth of customer data available to develop predictive models capable of anticipating churn behavior.

Rajendran et al. (2023) conducted a comprehensive study exploring machine learning approaches for churn prediction, underscoring the importance of proactive identification of at-risk customers. By harnessing techniques such as logistic regression, their work demonstrated promising results across various industry sectors, including telecommunications. Building upon this foundation, Arivazhagan and Subramanian (2020) introduced advancements in logistic regression techniques, incorporating regularization and optimization methods to enhance predictive accuracy. Their findings underscored the critical role of data preprocessing and algorithmic refinement in improving model performance, particularly in sectors characterized by high data volume and complexity.

The scale and diversity of telecom data present unique challenges and opportunities for predictive analytics. Ahmad, Jafar, and Aljoumaa (2019) delved into the realm of big data platforms, showcasing the potential of machine learning in handling vast datasets for churn prediction. Their work highlighted the need for scalable and efficient algorithms capable of processing real-time data streams, enabling timely intervention and personalized retention efforts. Furthermore, Xu, Ma, and Kim (2021) introduced innovative ensemble learning techniques, leveraging feature grouping to extract actionable insights from heterogeneous data sources. By integrating diverse models and leveraging feature engineering strategies, their approach demonstrated enhanced predictive power and robustness in churn prediction tasks.

The collective body of research in telecom churn prediction reflects a growing recognition of the transformative potential of machine learning in customer relationship management. From traditional logistic regression models to advanced ensemble methods, the landscape of predictive analytics continues to evolve, driven by a quest for greater accuracy, interpretability, and scalability. As telecom companies navigate the complexities of a rapidly evolving marketplace, leveraging these insights becomes imperative for sustaining competitiveness and fostering long-term customer loyalty.

III. Exploratory Data Analysis

A thorough Exploratory Data Analysis (EDA) was done on the dataset. The focus was on understanding the dataset's dynamics and how different factors relate to customer churn. The analysis reveals:
Data Distribution: Initial exploration indicates a significant imbalance with approximately 73% of customers not churning. This imbalance is crucial for predictive modeling, highlighting the need for strategies to mitigate skewness and avoid false negatives.
We further analyzed several factors affecting churn which includes tenure, contract type, gender distribution, senior-citizens, types of services used and the pricing, as shown in Figure 1 and Figure 2. Studies showed that the most important factors were the contract type and tenure. The dataset analysis highlights key factors influencing churn rates, such as the lack of online security and tech support services, the prevalence of month-to-month contracts, and higher monthly charges. To mitigate churn, strategies like offering enhanced service packages, promoting long-term contracts through incentives, and adopting flexible pricing strategies are recommended.

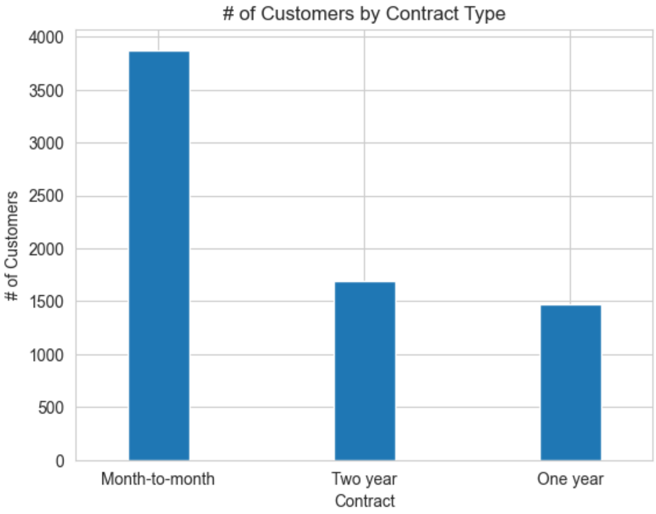

Fig. 1. Number of Customers by Contract Type

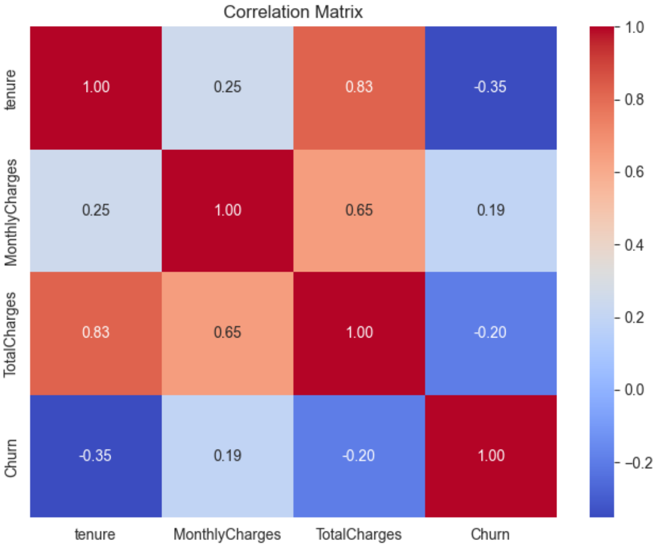

Fig. 2. Correlation Matrix of Churn

IV. Logistic Regression

Logistic regression was used to predict probability of customer churn. Probability of churn was calculated as

$$P(y=1|\mathbf{x}) = g(\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{xw}+b)}}$$

and

$$P(y=0|\mathbf{x}) = 1 - g(\mathbf{x})$$

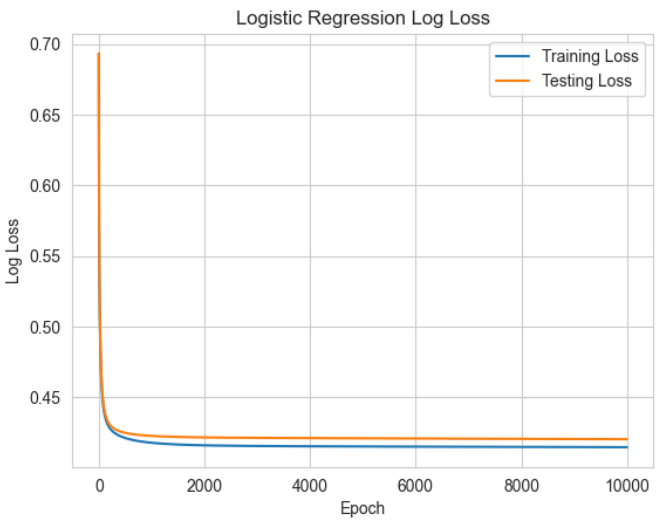

Fig. 3. Logistic Regression Log Loss

where y is churn, $\mathbf{x}$ is the data for that observation, and $\mathbf{w}$ and b are the weights and bias, respectively. The probability P will be $0 \leq P(y = 1|\mathbf{x}) \leq 1$.

The data was preprocessed prior to running the algorithm. All input data was scaled to be numeric in the range $\{0, 1\}$. $2/3$ of the data was used for training, with the rest reserved for validation.

The objective function used to optimize the gradients was the log loss function, calculated as

$$J = -(y \ln \hat{y} + (1 - y) \ln 1 - \hat{y})$$

where $\hat{y}$ is the predicted outcome from $g(\mathbf{x}$. Gradients of the log loss function were calculated as

$$\frac{\partial J}{\partial w_j} = (\hat{y} - y)\mathbf{x}_j$$

The model achieved a validation accuracy of 80.41%, as shown in Figure 3. The learning rate used was $\eta = 0.1$ and the model was run over 10,000 epochs.

- Epochs: 10000, $\eta : 0.1$
- Testing Accuracy: 80.41%, Precision: 67.19%, Recall: 52.73%, F1 Score: 59.09%

V. Decision Tree

A decision tree was implemented to predict churn using binary features. Numerical features including TotalCharges, MonthlyCharges, and Tenure were converted to 1's and 0's to represent whether the entry was above or below the feature mean. Boolean features including Partner, Dependents, and PhoneService were converted from $\{$Yes, No$\}$ to $\{1, 0\}$. Categorical features including OnlineBackup, StreamingTV, and PaymentMethod were each converted into multiple binary features to represent categories. For example, feature OnlineBackup had categories $\{$Yes, No, No internet service$\}$. It was converted to three features OnlineBackup_Yes, OnlineBackup_No, and OnlineBackup_No internet service. This preprocessing resulted in a total of 41 features.

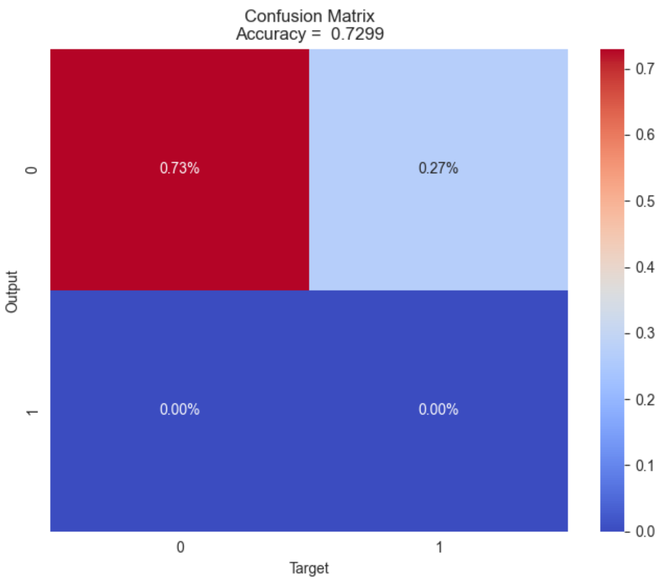

Fig. 4. Decision Tree Confusion Matrix

Preliminary implementation of the decision tree had an accuracy of 72.05%, as shown in Figure 4. Examination of the confusion matrix and metrics revealed that there were a significant number of false positives and false negatives.

To improve accuracy, features with the highest entropy were removed from the dataset. Machine learning methods for feature selection can include feature clustering, greedy feature selection, calculating entropy, calculating feature diversity, and other methods. [16] [17] For this application, entropy was calculated as

$$H(P(\nu_1)), ..., P(\nu_K)) = \sum_{i=1}^{K} (-(\nu_1)), ..., P(\nu_K) \log_K (P(\nu_i))$$

Removing the 20 features with highest entropy resulted in a validation accuracy of 78.18%, as shown in Figure 5.

Further feature reduction resulted in a validation accuracy of 79.16%, as shown in Figure 6. This improved accuracy demonstrates how feature reduction can improve model performance. Other methods for optimizing feature selection could result in further improved accuracy. [5]

VI. Random Forest

The Random Forest algorithm was also implemented for predicting churn. The model builds multiple decision trees and merges them together to get a more accurate and stable prediction. By combining the predictions of several trees, it reduces the risk of overfitting associated with individual decision trees. Preliminary analysis used the voting equation

$$P(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t$$

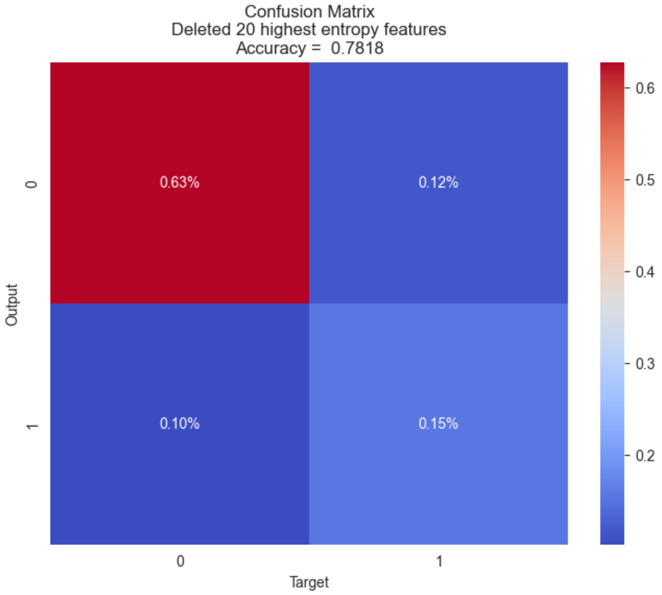

Fig. 5. Decision Tree Confusion Matrix: 20 Features

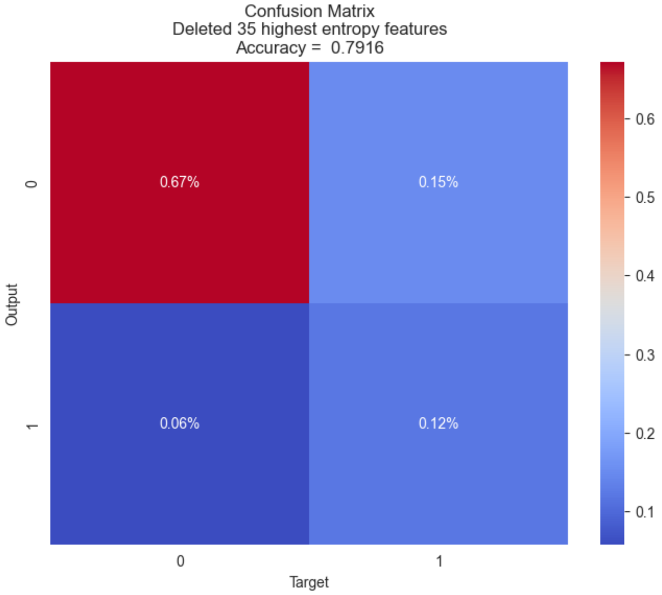

Fig. 6. Decision Tree Confusion Matrix: 6 Features

where P is the probability of a positive class. The predicted class is then

$$\hat{y} = (P(x) \geq 0.5)$$

By learning from the patterns in the dataset, the model aims to identify which factors contribute most significantly to customer decisions to leave the company. For example, features such as 'tenure' might indicate loyalty and satisfaction, whereas high 'MonthlyCharges' could contribute to customer dissatisfaction and churn. The model's training process involves adjusting to the complexities of the dataset, such as handling both numerical and categorical variables,

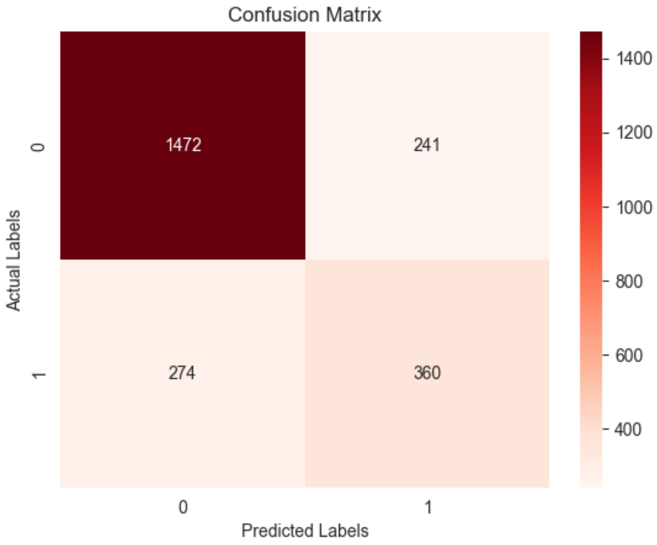

Fig. 7. Random Forest Confusion Matrix; Trees: 100, Max Depth: 41

and interpreting the interactions between different customer attributes.We applied the RandomForest algorithm to predict customer churn based on telecom data. Through this analysis, we obtained three notable results:

1st result: fig7

- Trees: 100, Max Depth: 41
- Training Accuracy: 77.39%, Precision: 57.84%, Recall: 51.66%, F1 Score: 54.58%
- Testing Accuracy: 78.06%, Precision: 59.90%, Recall: 56.78%, F1 Score: 58.30%
- Overall Accuracy: 78.06%

This configuration demonstrates a balanced model with a depth allowing for complex patterns to be recognized, contributing to its relatively high performance. However, the model was predicting negative results far more frequently that positive results, causing the skewed results seen in Figure 7. To attempt to correct for this, the model was increased to 500 trees, as shown in Figure 8.

2nd result: fig 8

- Trees: 500, Max Depth: 41
- Training Accuracy: 75.28%, Precision: 72.02%, Recall: 9.80%, F1 Score: 17.25%
- Testing Accuracy: 74.78%, Precision: 71.43%, Recall: 11.04%, F1 Score: 19.13%
- Overall Accuracy: 74.78%

Increasing the number of trees to 500 did not significantly improve the model's predictive power for this dataset, highlighting a possible overfit on the training data.

3rd result: fig 9

- Trees: 100, Max Depth: 4
- Training Accuracy: 77.04%, Precision: 56.53%, Recall: 55.06%, F1 Score: 55.78%
- Testing Accuracy: 77.89%, Precision: 58.97%, Recall: 59.62%, F1 Score: 59.29%

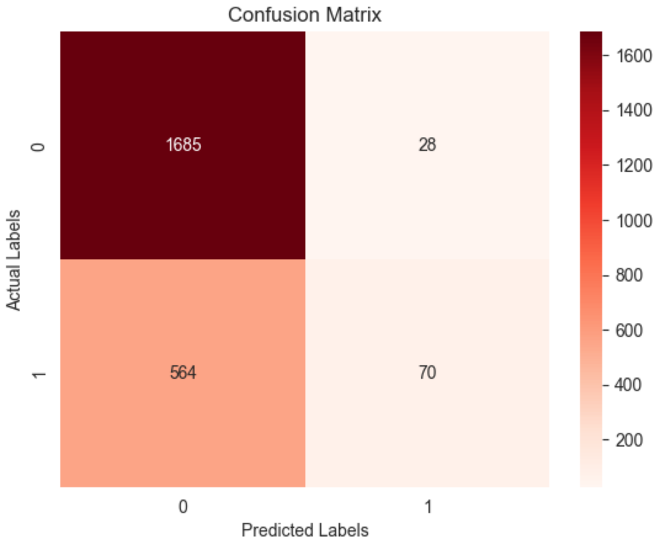

Fig. 8. Random Forest Confusion Matrix; Trees: 500, Max Depth: 41

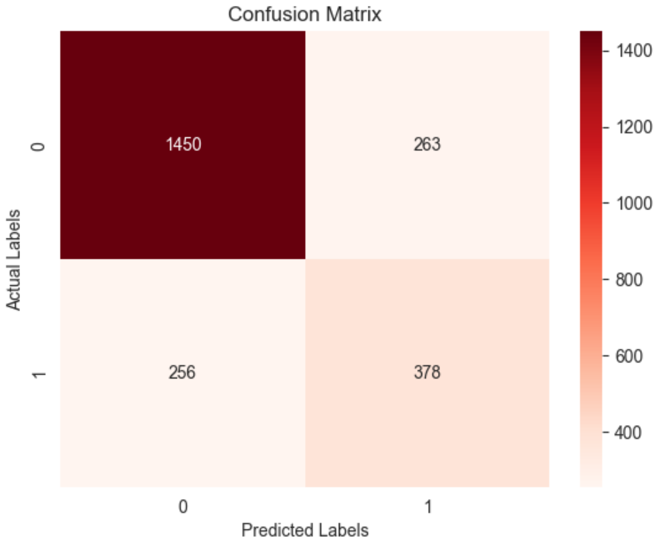

Fig. 9. Random Forest Confusion Matrix; Trees: 100, Max Depth: 4

- Overall Accuracy: 77.89%

This third model configuration, with a shallower depth, suggests an improvement in generalization over the testing set, offering a robust predictive capability while mitigating overfitting risks. These results were achieved by training the RandomForest model with varying numbers of trees and maximum depth parameters, evaluating its performance on both training and testing datasets to assess accuracy, precision, recall, and the F1 score. This analysis highlights the importance of tuning model parameters to achieve optimal performance in machine learning tasks.

Additionally, we refined the RandomForest model with 500 trees, a maximum depth of 21, and utilizing the 21 lowest-entropy features in the dataset. This feature reduction was effective in the single decision tree models. The outcomes were as follows:

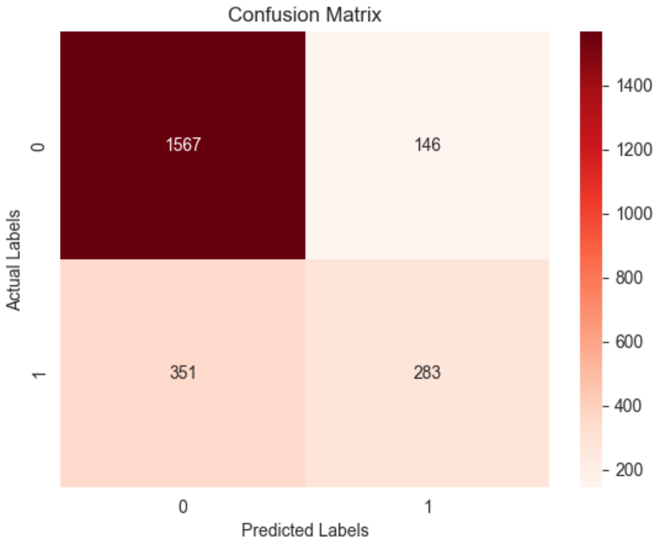

Fig. 10. Random Forest Confusion Matrix; Trees: 100, Max Depth: 21, Features: 21

Additional result: fig 10

- Trees: 500, Max Depth: 21, Features: 21
- Training Accuracy: 78.49%, Precision: 64.19%, Recall: 41.21%, F1 Score: 50.20%
- Testing Accuracy: 78.82%, Precision: 65.97%, Recall: 44.64%, F1 Score: 53.25%
- Overall Accuracy: 78.82%

This configuration underscores the effect of leveraging the full range of features available in the dataset. By using the 21 lowest-entropy features, the model is able to capture more nuanced patterns and interactions among variables, leading to a notable improvement in predictive accuracy and precision. The balanced depth and number of trees also contribute to an enhanced ability to generalize across the testing set, showcasing a comprehensive approach to understanding and mitigating customer churn.

VII. Ensemble

An Ensemble model was created to integrate the predictive capabilities of logistic regression and random forest. The logistic regression provides probabilities, while the random forest outputs are converted to match this format, and a final prediction is made by averaging the predictions from both models.

$$\text{Prediction} = \frac{\text{LR Probability} + \text{RF Prediction}}{2}$$

It produced the following results:

- Accuracy: 78.69%
- Precision: 61.44%
- Recall: 58.01%
- F1 Score: 59.68%

A confusion matrix was also generated for the same. This indicates that the combined model correctly predicts churn in nearly four out of five cases. These metrics collectively suggest that the ensemble approach is effective, though there may be room for further optimization.

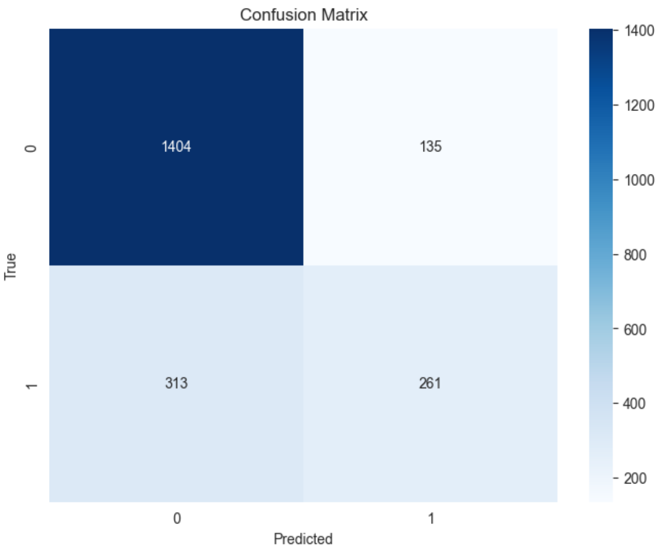

Fig. 11. Ensemble Confusion Matrix

VIII. Observations and Comparison

Performance Across Models:

Random Forest showcased the highest recall and F1 Score, demonstrating its superior ability to identify churn customers accurately without excessively misclassifying non-churn customers. This model is adept at handling the complexity of the dataset and avoiding overfitting, which is a common pitfall of Decision Trees.

Logistic Regression had the highest accuracy and showed commendable performance across all metrics. Though, Random Forest had lower accuracy, it performed better overall. This model's strength lies in its simplicity and interpretability, which can be particularly useful when it is important to understand the drivers of churn.

The Decision Tree presented a straightforward model but with lower performance metrics than the Random Forest. It remains valuable for its interpretability and ease of implementation.

The Ensemble model did not exceed the performance of Random Forest or Logistic Regression. This indicates that the way the models were combined may not have been the most effective. The ensemble method's precision was notably high, but its recall was lacking, suggesting it may not identify as many actual churn cases as Random Forest.

Precision and Recall Trade-off: Despite the high precision of the Ensemble model, its low recall suggests it is less suitable for scenarios where it is critical to capture as many true churn cases as possible. Random Forest, with its balanced recall and precision reflected in the F1 Score, would be preferable in scenarios where both identifying churners and avoiding false churn predictions are important.

F1 Score: Random Forest achieved the highest F1 Score, indicating it maintained a balance between recall and precision. This suggests that Random Forest is the most suitable model for our churn prediction task since it captures the harmonic mean of precision and recall, essential for our dataset.

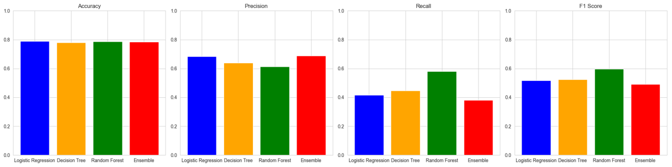

Fig. 12. Comparison between models

IX. Conclusions and Analysis

The evaluation of our churn prediction models reveals that the Random Forest algorithm offers the most balanced performance, with the highest F1 Score indicating its proficiency in accurately predicting churn. Despite Logistic Regression's high precision and interpretability, which provides clear insights for strategic decision-making, it falls short in achieving the best balance between precision and recall.

The Ensemble model did not outperform its individual counterparts, suggesting that more advanced techniques like weighted voting or stacking might be needed to fully capitalize on the strengths of different models.

The study also notes that feature quality significantly impacts model performance. Specifically, noisy and high-entropy features reduced the efficacy of both the Logistic Regression and Decision Tree models. Through entropy-based feature selection and meticulous parameter tuning, improvements were noted, although this process proved to be time-consuming.

Considering the complexities and financial implications of customer retention, our findings suggest adopting the Random Forest model as it aligns well with the need for an accurate and reliable churn prediction model. Continuous refinement and adaptation of this model are recommended to ensure its efficacy as customer behaviors and market conditions evolve.

Moreover, model performance should be regularly validated with new data and across different customer segments to maintain a robust churn prediction system. The effectiveness of any model may fluctuate based on dataset characteristics and demographic variances, underlining the necessity for persistent testing and adaptation in our churn prediction strategy.

X. Future Work

For future advancements in churn prediction, efforts will focus on:

Feature Optimization: Refining feature selection to better handle noisy data and exploring feature engineering for new insights.

Ensemble Techniques: Investigating advanced ensemble methods such as boosting and stacking to improve predictive performance.

Hyperparameter Refinement: More rigorous tuning of hyperparameters, particularly for the Random Forest model.

Class Imbalance Solutions: Implementing methods to address class imbalance, such as synthetic data generation or algorithmic adjustments.

Continuous Model Improvement: Regularly updating models with the latest data and validating across diverse customer groups to ensure accuracy and generalizability.

Exploring AI Trends: Incorporating new developments in AI to enhance the churn prediction models' capabilities.

By concentrating on these areas, we aim to build more accurate and robust models for predicting customer churn.

References

[1] Stoltzfus, J.C. (2011), Logistic Regression: A Brief Primer. Academic Emergency Medicine, 18: 1099-1104. https://doi-org.ezproxy2.library.drexel.edu/10.1111/j.1553-2712.2011.01185.x

[2] M. Sergue, ?Customer Churn Analysis and Prediction using Machine Learning for a B2B SaaS company,? Dissertation, 2020. https://www.diva-portal.org/smash/get/diva2:1426161/FULLTEXT01.pdf

[3] A. Rudalv, ?Predicting Customer Churn in E-Commerce Using Statistical Modeling and Feature Importance Analysis A Comparison of Random Forest and Logistic Regression Approaches.? Available: https://umu.diva-portal.org/smash/get/diva2:1781292/FULLTEXT01.pdf

[4] Rajendran, Srinivasan & Devarajan, Rajeswari & Elangovan, G.. (2023). Customer Churn Prediction Using Machine Learning Approaches. 1-6. 10.1109/ICECONF57129.2023.10083813.

[5] B. Arivazhagan & R. S. Sankara Subramanian, ?Customer Churn Prediction using Logistic Regression with Regularization and Optimization Technique,? Regular, vol. 9, no. 9, pp. 334?339, Jul. 2020, doi: https://doi.org/10.35940/ijitee.i7219.079920.

[6] M. Manasa, ?Telecom Customer Churn Prediction,? International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 5, pp. 2857?2862, May 2020, doi: https://doi.org/10.22214/ijraset.2020.5479.

[7] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data 6, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

[8] Sana JK, Abedin MZ, Rahman MS, Rahman MS (2022) A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. PLoS ONE 17(12): e0278095. https://doi.org/10.1371/journal.pone.0278095

[9] Richter, Yossi & Yom-Tov, Elad & Slonim, Noam. (2010). Predicting Customer Churn in Mobile Networks through Analysis of Social Groups. Proceedings of the 10th SIAM International Conference on Data Mining, SDM 2010. 732-741. 10.1137/1.9781611972801.64.

[10] H. Wu, A High-Performance Customer Churn Prediction System based on Self-Attention,?arXiv.org, Jun. 03, 2022. https://arxiv.org/abs/2206.01523 (accessed Mar. 16, 2024).

[11] K. Coussement and D. Van den Poel, ?Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques,? Expert Systems with Applications, vol. 34, no. 1, pp. 313?327, Jan. 2008, doi: https://doi.org/10.1016/j.eswa.2006.09.038.

[12] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, ?Customer Churn Prediction in Telecom Sector using Machine Learning Techniques,? Results in Control and Optimization, p. 100342, Nov. 2023, doi: https://doi.org/10.1016/j.rico.2023.100342.

[13] Xu T, Ma Y, Kim K. Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. Applied Sciences. 2021; 11(11):4742. https://doi.org/10.3390/app11114742

[14] A. Navada, A. N. Ansari, S. Patil and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," 2011. IEEE Control and System Graduate Research Colloquium, Shah Alam, Malaysia, 2011, pp. 37-42, doi: 10.1109/ICSGRC.2011.5991826.

[15] Monard, Maria-Carolina & Batista, Gustavo. (2002). Learning with skewed class distrihutions. Adv. Log. Artif. Intell. Robot. LAPTEC. 2002. https://www.researchgate.net/publication/311396259 (accessed Mar. 16, 2024).

[16] R. A. Janik, ?Entropy from Machine Learning,? arXiv.org, Oct. 24, 2019. https://arxiv.org/abs/1909.10831 (accessed Mar. 16, 2024).

[17] J. Cai, J. Luo, S. Wang, and S. Yang, ?Feature selection in machine learning: A new perspective,? Neurocomputing, vol. 300, pp. 70?79, Jul. 2018, doi: https://doi.org/10.1016/j.neucom.2017.11.077.