Customer Churn Predictive Modeling

William Sparks
Eastern Kentucky University

william_sparks7@mymail.eku.edu

***Abstract*—Customer churn is a significant challenge faced by businesses, particularly in subscription-based industries where customer retention is essential for sustaining revenue and reducing acquisition costs. In this project, we develop a predictive model aimed at identifying customers at risk of churning. Our initial approach involves training a Random Forest model which achieves moderate success, with an accuracy of approximately 80%. However, this base model struggles with identifying churners accurately. To address this, we implement strategies such as feature selection and decision threshold optimization. These enhancements improve the model's F1-score for the churn class, illustrating the importance of tuning parameters when dealing with imbalanced datasets. This project underscores the necessity of balancing false positives and false negatives in predictive modeling and presents a suggestion for further research in customer retention.**

***Index Terms*—Customer Churn, Random Forest, Predictive Modeling, Feature Selection, F1-score, ROC-AUC, Decision Threshold Optimization**

I. Introduction

Customer retention is a critical challenge for modern companies, particularly in subscription-based industries like telecommunications, banking, and software-as-a-service (SaaS). Retaining customers is essential for maintaining revenue and profitability, as churn—the loss of clients or subscribers—significantly disrupts a company's financial stability. Research shows that acquiring a new customer is substantially more expensive than retaining an existing one, with estimates suggesting that acquisition costs are five to ten times higher. Moreover, a 5% increase in retention boosts profits by as much as 25% to 95%, depending on the industry [1]. As a result, predictive models for identifying customers at risk of churn have become vital tools for improving retention rates.

Predictive churn models analyze historical customer data to anticipate which customers are likely to leave. These models rely on a variety of customer attributes such as demographics, purchase behavior, service usage, and engagement. Since churn prediction is often an imbalanced classification problem—where the number of churners is significantly smaller than non-churners—machine learning models must handle this imbalance to avoid biased results.

Random Forest, an ensemble learning algorithm, is particularly well-suited for churn prediction due to its ability to manage large datasets and avoid overfitting. It also handles class imbalances effectively by adjusting for the minority class, making it a reliable choice for predicting customer churn. By combining predictions from multiple decision trees, Random Forest achieves more accurate and stable results compared to simpler models. See Fig.1

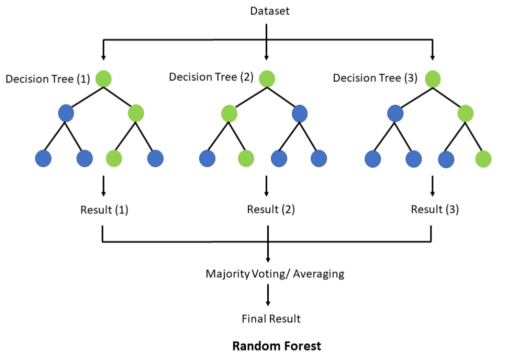

Fig. 1. Random Forest Schema [2]

In this project, we train a Random Forest model to predict customer churn and achieved an initial accuracy of 80%. However, the model underperforms in identifying churners: a common issue in churn prediction due to the imbalanced nature of the dataset. To address this, we implement feature selection and decision threshold optimization. Feature selection helps improve generality by focusing on the most important predictors, while lowering the decision threshold increased the model's recall, allowing it to identify more churners. Although this adjustment leads to more false positives, it proves beneficial in a business context where missing churners is more costly than misclassified non-churners.

Our findings demonstrate that tuning key model parameters, such as the decision threshold, can significantly improve a model's ability to detect churners in imbalanced datasets. These adjustments are crucial for businesses aiming to enhance customer retention through predictive modeling. This project highlights the advantages of using Random Forest for churn prediction and provides a foundation for future research on more advanced models and techniques, such as Gradient Boosting and Neural Networks, to further improve performance.

II. Related Work

The problem of customer churn prediction has been extensively studied across various industries, including telecommunications, banking, retail, and software-as-a-service (SaaS). Numerous approaches have been explored, ranging from traditional statistical models to advanced machine learning techniques. The simplest approach often involves Logistic Regression, a widely used statistical model for binary classification problems. Logistic Regression is known for its ease of interpretation and efficiency in processing structured data. However, while Logistic Regression models provide valuable insights into feature importance, its inability to capture complex, non-linear relationships between variables can limit its effectiveness in predictive modeling [3].

To overcome the limitations of traditional models, more sophisticated machine learning techniques have been introduced, such as Decision Trees, Gradient Boosting, and Neural Networks. For instance, Random Forest, an ensemble method based on decision trees, has gained popularity due to its ability to handle high-dimensional data and reduce overfitting [4].

Gradient Boosting, another powerful ensemble technique, has been shown to outperform single decision tree models by iteratively correcting the mistakes of weak learners. In the telecommunications industry, [5] applied a Logistic Regression model and reported an accuracy of 79%. However, Gradient Boosting methods, such as XGBoost, have consistently been found to deliver higher accuracy in customer churn prediction by capturing complex, non-linear patterns that Logistic Regression often misses [6]. Gradient Boosting has also been effective in handling class imbalance, a common issue in churn prediction where the number of churners is typically far smaller than non-churners [7].

In addition to Gradient Boosting, Neural Networks have gained traction in recent years due to their ability to model complex, high-dimensional datasets. [8] explored the use of Gradient Boosting combined with a recurrent neural network (RNN) to predict churn in customer-centric environments. Their findings indicated that the combination of these two techniques led to improved accuracy, especially in detecting churners within imbalanced datasets. This hybrid approach demonstrates the competitive edge of combining advanced machine learning algorithms to improve predictive power.

Reinforcement learning has also been explored as a method to not only predict churn but to optimize customer retention strategies dynamically. According to [9], reinforcement learning models can be used to recommend personalized actions aimed at reducing churn, such as offering discounts or targeted customer support. These models differ from traditional predictive models by continuously learning from customer interactions, thus improving their ability to make real-time decisions.

Another emerging area is the use of Natural Language Processing (NLP) to enhance churn prediction. By analyzing customer reviews, feedback, and unstructured data such as call logs, NLP techniques can provide additional features that improve the accuracy of predictive models [10]. This is particularly useful in industries where customer sentiment and engagement are critical indicators of churn.

In summary, the research on customer churn prediction has evolved from simple statistical models to advanced machine learning and deep learning techniques. While Logistic Regression remains a useful baseline model, methods like Random Forest, Gradient Boosting, and Neural Networks have proven to be more effective at capturing complex patterns in large datasets. Emerging techniques like reinforcement learning and NLP offer promising directions for future research, providing businesses with more dynamic and actionable insights for reducing customer churn.

III. Methodology

Our approach consists of several key steps, including data preprocessing, feature selection, model training, and threshold optimization. Each of these steps plays a critical role in the overall performance of the model.

A. Data Preprocessing

The dataset [11] used in this project contains over 7,000 customer records, with features ranging from demographic information (e.g., age, gender) to service usage metrics (e.g., number of active subscriptions, customer engagement levels). Before training the model, we applied several preprocessing steps to ensure data quality.

There were initially 11 missing values for TotalCharges due to being inconvertible to numerical type under a transformation. Since this is relatively small in comparison to the number of records, we decided to drop these rows. Categorical features, such as customer location and subscription type, were encoded using one-hot encoding to convert them into numerical form. This ensures that the machine learning algorithm can process the data effectively. Additionally, to address the issue of class imbalance (since only a small proportion of customers churned), we chose to stratify the train test split data so that our sets retain the underlying class distribution of the dataset. We also set the class_weight parameter in RandomForestClassifier to 'balanced' to automatically adjust the weights and increase the weight for the minority class.

B. Feature Selection

Feature selection is essential for improving the model's performance and for generalization. By focusing on the most relevant features, we reduce noise and prevent overfitting. We used the feature importance scores generated by the Random Forest model to rank the features by their predictive value. Features with low importance were removed (defined as importance $< \delta = 0.02$), allowing us to retain 12 of the 27 original features (feature set after one-hot encoding). We kept the 12 most important features in Fig. 2. This step only saw a slight decrease in accuracy of 4% but also simplified the model, making it easier to interpret the results and implement actionable insights for customer retention strategies.

C. Model Training and Evaluation

The Random Forest model was trained using a 80-20 train-test split. We opted for 100 estimators which struck a balance between model complexity and performance. We performed hyperparameter tuning by iteratively testing decision thresholds in an effort to maximize accuracy scores for the churn class. Special emphasis was placed on recall and F1-score, given the importance of correctly identifying churners.

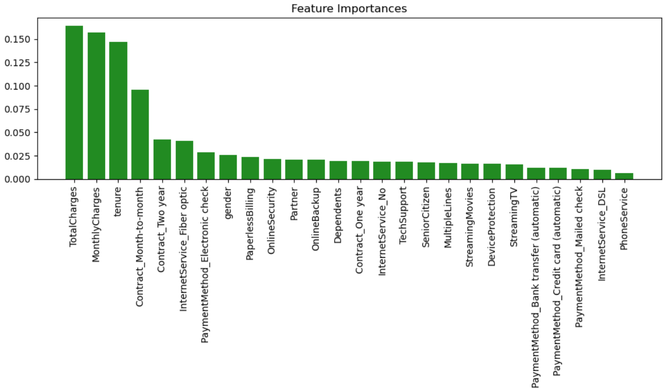

Fig. 2. Feature Importances

The model was evaluated using the following metrics:

- Accuracy: Measures the overall correctness of the model.
- Precision: Measures the proportion of predicted churners that were actually churners.
- Recall: Measures the proportion of actual churners that were correctly predicted by the model.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance on churners.

D. *Threshold Optimization*

One of the critical aspects of churn prediction is the trade-off between precision and recall. In our initial model, the default decision threshold of 0.5 resulted in high precision but low recall, meaning that many churners were misclassified as non-churners. This could be costly for businesses, as failing to identify churners may result in lost customers.

To improve recall, we optimize the decision threshold. By lowering the threshold to 0.31, we are able to increase the recall of the model significantly, identifying more churners at the cost of a higher false positive rate. This trade-off is often acceptable in business contexts, where it is less costly to intervene with customers who may not churn than to lose customers who were not predicted to leave. A Precision-Recall vs. Threshold plot appears in Fig. 3 below.

IV. Results

The initial Random Forest model achieved an accuracy of approximately 80%, but it struggled to identify churners effectively. After implementing feature selection and threshold optimization, the model's performance improved across several key metrics, as shown below:

- Accuracy: 75%
- Precision (Churn): 53%
- Recall (Churn): **71%**
- F1-score (Churn): 60%
- ROC-AUC: 0.8

These results demonstrate the importance of threshold optimization in improving the model's ability to detect churners.

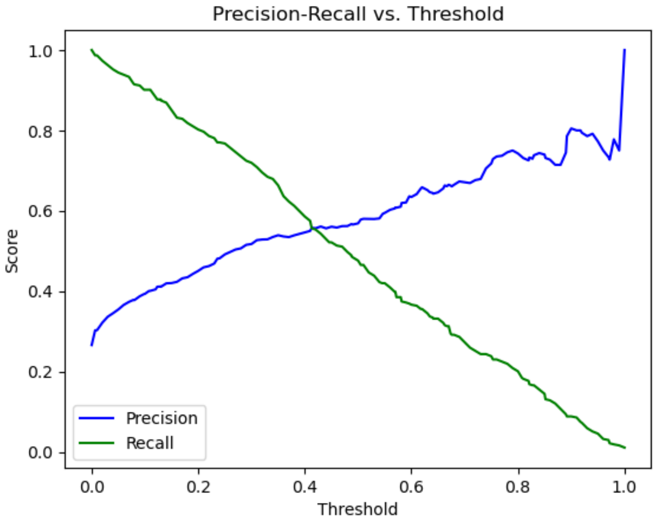

Fig. 3. Precision-Recall Curve

Fig. 4 shows the ROC curve for the final model, with an AUC of 0.8, indicating a fairly strong discriminatory ability.

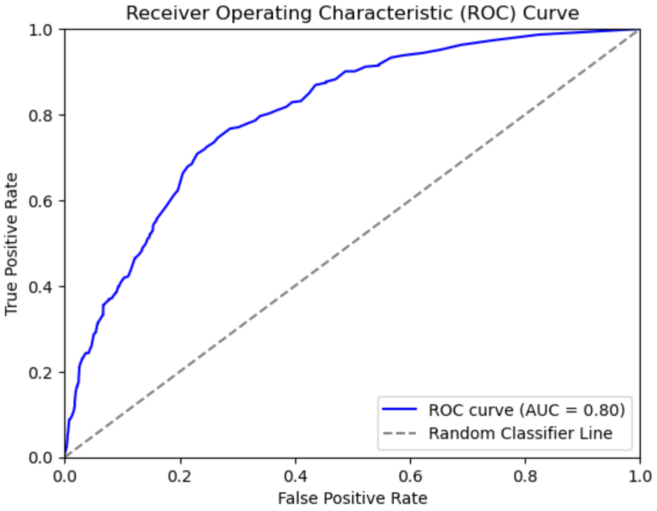

Fig. 4. ROC Curve of Random Forest Model

This was a significant improvement for the class 1 results when compared to the metrics of the original model, prior to dropping features and fine-tuning decision threshold:

- Accuracy: 79%
- Precision (Churn): 64%
- Recall (Churn): **49%**
- F1-score (Churn): 55%
- ROC-AUC: 0.82

Although certain metrics decreased after simplifying and fine-tuning the model, this is acceptable since the recall is the most important of these scores. We observed a 45% boost in recall for the churn class. The base model performs exceptionally well with respect to predicting the non-churn class, but this is not the primary area of interest for this application.

A. Confusion Matrix

Table I provides the confusion matrix for the final model, illustrating the trade-offs between false positives and false negatives. The model correctly classified 85% of churners, but this came at the expense of a higher false positive rate.

TABLE I
CONFUSION MATRIX

	Predicted Non-Churn	Predicted Churn
Actual Non-Churn	906	127
Actual Churn	200	174

V. CONCLUSION

This project demonstrates the effectiveness of Random Forest models for customer churn prediction, particularly when combined with feature selection and threshold optimization techniques. Random Forest, being an ensemble learning method, proved highly robust in handling large datasets and managing class imbalances. By adjusting the decision threshold, we improved the model's recall, making it more effective at identifying churners, even at the cost of a higher false positive rate. This trade-off is often acceptable in business contexts where the cost of losing a valuable customer outweighs the expense of offering retention incentives to those less likely to churn.

Our findings suggest that businesses can benefit significantly from targeted retention efforts based on model predictions. Companies in industries like telecommunications, banking, and SaaS could use these insights to prioritize intervention efforts, focusing on high-risk customers. Given the feature importances, such as total charges, tenure, and contract type, businesses should analyze long-term billing history and customer engagement metrics closely. These insights can help refine customer segmentation and develop more personalized retention campaigns.

A key retention strategy is offering incentives to customers nearing contract expiration or exhibiting behaviors linked to churn. Discounts, service upgrades, or loyalty rewards can increase the opportunity cost of leaving, reducing churn likelihood. For long-tenure customers, offering rate reductions or personalized services can maintain loyalty and avoid the loss of high-value customers.

Predictive models also enable proactive engagement by addressing potential pain points before they lead to churn. For example, companies can track feature usage or customer complaints to intervene with personalized solutions. This kind of proactive strategy is essential for retaining customers in competitive markets.

In future work, we plan to explore more advanced machine learning models such as XGBoost, CatBoost, and LightGBM, which often outperform Random Forest in churn prediction tasks. These models handle non-linear relationships in data more effectively. Additionally, investigating hyperparameter optimization techniques such as grid search, random search, and Bayesian optimization will help improve model performance. Utilizing cross-validation during tuning can ensure that models generalize well to new data.

We will also explore incorporating additional data sources, such as behavioral metrics from web usage or social media interactions. Using Natural Language Processing (NLP) to analyze customer feedback could provide more insights into customer sentiment, further refining retention strategies.

Lastly, real-time deployment of churn prediction models presents a valuable future direction. Developing models that continuously update predictions could enable businesses to intervene before churn occurs. Integration with customer relationship management (CRM) systems could automate personalized retention efforts.

In conclusion, while Random Forest models are effective for churn prediction, there is potential for improvement by exploring advanced algorithms, fine-tuning hyperparameters, and integrating additional data. This will help businesses retain customers more effectively and improve long-term profitability.

REFERENCES

[1] Propello Cloud, "The Benefits of Customer Retention," *Propello Cloud Blog*, 2024. [Online]. Available: https://blog.propellocloud.com/benefits-of-customer-retention.

[2] Random forest explain.png, Wikimedia Commons. Available at: https://commons.wikimedia.org/wiki/File:Random_forest_explain.png [Accessed October 9, 2024].

[3] M. A. Shaikhsurab and P. Magadum, "Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach," arXiv preprint arXiv:2408.16284, 2024. [Online]. Available: https://arxiv.org/abs/2408.16284.

[4] "Random Forests: A Guide," Institute of Data. [Online]. Available: https://www.institutedata.com/blog/random-forests-a-guide/

[5] A. Sundararajan and K. Gursoy, "Telecom Customer Churn Prediction," *Rutgers University Libraries*, 2020. [Online]. Available: https://doi.org/10.7282/t3-76xm-de75.

[6] R. Dwidarma, S. D. Permai, and J. Harefa, "Comparison of Logistic Regression and XGBoost for Predicting Potential Debtors," in *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, 2021, pp. 1-6. [Online]. Available: https://ieeexplore.ieee.org/document/9574350.

[7] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Scientific Reports*, vol. 13, no. 1, p. 14441, Sep. 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37660198/.

[8] S. Haddadi, A. Farshidvard, F. dos Santos Silva, J. dos Reis, and M. da Silva Reis, "Customer Churn Prediction in Imbalanced Datasets with Resampling Methods: A Comparative Study," *Expert Systems with Applications*, vol. 246, 2024, 123086. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.123086.

[9] T. Khan, S. A. Idrisi, H. Patil, and J. Dongradive, "Customer Churn Analysis Using Deep Reinforcement Learning Approach," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 6, no. 1, pp. 1-6, Jan.-Feb. 2024. [Online]. Available: https://www.ijfmr.com/papers/2024/1/11712.pdf.

[10] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," *Knowledge-Based Systems*, vol. 212, p. 106586, 2021. [Online]. Available: https://doi.org/10.1016/j.knosys.2020.106586.

[11] "Telco Customer Churn Dataset," Kaggle. Available at: https://www.kaggle.com/datasets/blastchar/telco-customer-churn [Accessed October 9, 2024].