JANOLI International Journals of Data Science (JIJDS)

Volume. 1, Issue 2, February 2025

Leveraging Ensemble Learning and Feature Engineering for Enhanced Predictive Accuracy in Customer Churn Prediction

Authors:

Pradeep Upadhyay, Niet, NIMS UNIVERSITY, pradeep.upadhyay@nimsuniversity.org

Keywords:

Customer Churn, Ensemble Learning, Feature Engineering, Predictive Modeling, Machine Learning, Classification, Data Science, Customer Relationship Management, Model Optimization, Imbalanced Data

Article History:

Received: 03 February 2025; Revised: 10 February 2025; Accepted: 17 February 2025; Published: 28 February 2025

Abstract:

Customer churn prediction is a critical challenge for businesses seeking to maintain and grow their customer base. This research investigates the application of ensemble learning techniques combined with advanced feature engineering to enhance the accuracy of churn prediction models. We explore several ensemble methods, including Random Forest, Gradient Boosting Machines (GBM), and XGBoost, and evaluate their performance against traditional machine learning algorithms. Furthermore, we implement a comprehensive feature engineering strategy, incorporating techniques such as interaction feature generation, polynomial features, and domain-specific feature extraction. Our results demonstrate that the proposed approach significantly improves churn prediction accuracy compared to baseline models, offering valuable insights for customer retention strategies. The study highlights the importance of both model selection and feature engineering in building robust and effective churn prediction systems.

1. Introduction

In today's competitive business landscape, customer retention is paramount. Acquiring new customers is often significantly more expensive than retaining existing ones, making customer churn prediction a crucial task for businesses across various industries. Customer churn, also known as customer attrition, refers to the phenomenon where customers cease their relationship with a company. Accurately predicting which customers are likely to churn allows businesses to proactively implement targeted retention strategies, minimizing revenue loss and maximizing customer lifetime value.

Traditional approaches to churn prediction often rely on simple statistical models or basic machine learning algorithms. However, these methods may struggle to capture the complex and non-linear relationships present in customer data. The performance of these models is also heavily dependent on the quality and relevance of the features used.

This research aims to address these limitations by exploring the application of advanced ensemble learning techniques coupled with comprehensive feature engineering strategies for enhanced churn prediction. Ensemble learning combines multiple base models to create a stronger, more robust predictive model. Feature engineering involves transforming raw data into meaningful features that can improve the performance of machine learning algorithms.

Problem Statement:

The existing methods for churn prediction often suffer from limited accuracy due to the complexity of customer behavior and the limitations of traditional machine learning algorithms and feature sets. There is a need for more sophisticated approaches that can effectively capture the intricate patterns and relationships within customer data to improve prediction accuracy and provide actionable insights.

Objectives:

The primary objectives of this research are:

To evaluate the performance of various ensemble learning techniques (Random Forest, GBM, XGBoost) for customer churn prediction.

To develop and implement a comprehensive feature engineering strategy to enhance the predictive power of churn prediction models.

To compare the performance of ensemble learning models with feature engineering against traditional machine learning algorithms.

To identify the most important features for churn prediction and provide insights for customer retention strategies.

To build a robust and accurate churn prediction model that can be deployed in a real-world business environment.

2. Literature Review

Customer churn prediction has been extensively studied in the literature. Several studies have explored various machine learning techniques for predicting customer churn across different industries.

1. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

This paper emphasizes the importance of profit-driven churn prediction, incorporating the cost of retention campaigns into the model evaluation process. The authors compare different data mining techniques, including logistic regression and support vector machines, and demonstrate the economic benefits of churn prediction.

Strengths: Focuses on the business value of churn prediction. Weaknesses: Limited exploration of advanced machine learning techniques.

2. Coussement, K., Benoit, D. F., & Lessmann, S. (2017). A comparative study of data mining methods for customer churn prediction: Case study of a Belgian mobile telecom operator. European Journal of Operational Research, 260(2), 616-629.

This study compares the performance of various data mining methods, including decision trees, neural networks, and support vector machines, for churn prediction in the telecom industry. The authors find that neural networks generally outperform other methods.

Strengths: Comprehensive comparison of different data mining techniques. Weaknesses: Does not explore ensemble learning methods in depth.

3. Idris, A., Khan, A., & Lee, Y. (2012). Genetic programming based feature construction and ensemble learning for churn prediction. Expert Systems with Applications, 39(10), 7902-7910.

This paper proposes a genetic programming-based approach for feature construction and ensemble learning for churn prediction. The authors demonstrate that their approach can improve prediction accuracy compared to traditional methods.

Strengths: Innovative approach to feature engineering. Weaknesses: Computationally intensive.

4. Lin, C. H., & Wang, C. H. (2015). Applying data mining techniques to predict churn behavior of customers in the telecommunications industry. Expert Systems with Applications, 42(16), 6301-6308.

This study applies data mining techniques, including decision trees and neural networks, to predict churn behavior in the telecommunications industry. The authors identify key factors influencing churn and develop a churn prediction model.

Strengths: Focuses on the telecommunications industry. Weaknesses: Limited exploration of advanced feature engineering techniques.

5. Lessmann, S., Baesens, B., Seow, H. V., & Golden, B. L. (2015). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. IEEE Transactions on Software Engineering, 41(5), 455-472.

While focused on software defect prediction, this paper provides a valuable framework for benchmarking classification models, which can be applied to churn prediction. The authors emphasize the importance of using appropriate evaluation metrics and statistical significance tests.

Strengths: Provides a robust framework for model evaluation. Weaknesses: Not directly focused on churn prediction.

6. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(9), 11058-11066.

This paper addresses the issue of class imbalance in customer churn prediction, where the number of churned customers is significantly lower than the number of non-churned customers. The authors explore various techniques for handling class imbalance, including oversampling and undersampling.

Strengths: Addresses the important issue of class imbalance. Weaknesses: Limited exploration of advanced machine learning techniques.

7. Xie, X., Li, Q., & Xie, H. (2009). Data mining methods for customer churn prediction. International Conference on Management Science and Engineering, 2009.

This conference paper explores various data mining methods for customer churn prediction, including decision trees, neural networks, and support vector machines. The authors compare the performance of these methods and identify key factors influencing churn.

Strengths: Provides a broad overview of data mining methods for churn prediction. Weaknesses: Limited depth in the analysis of each method.

8. Ho, T. K. (1995). Random decision forests. Proceedings of the 3rd international conference on document analysis and recognition, 1, 278-282.

This seminal paper introduces the Random Forest algorithm, a powerful ensemble learning technique that has been widely used for classification and regression tasks, including churn prediction.

Strengths: Introduces a widely used and effective algorithm. Weaknesses: None, as it's a foundational paper.

9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794.

This paper introduces XGBoost, a highly optimized and scalable gradient boosting algorithm that has achieved state-of-the-art results in various machine learning competitions and applications, including churn prediction.

Strengths: Introduces a powerful and widely used algorithm. Weaknesses: Can be computationally expensive for very large datasets.

10. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

This paper provides a theoretical foundation for Gradient Boosting Machines (GBM), a powerful ensemble learning technique that combines multiple weak learners to create a strong predictive model.

Strengths: Provides a theoretical foundation for GBM. Weaknesses: Can be sensitive to overfitting.

Critical Analysis of Previous Work:

While previous studies have explored various machine learning techniques for churn prediction, several limitations remain. Many studies focus on comparing different algorithms without adequately addressing the importance of feature engineering. Furthermore, some studies do not explicitly address the issue of class imbalance, which is common in churn prediction datasets. Additionally, the generalizability of the findings may be limited due to the specific datasets and industries considered. This research aims to address these limitations by combining advanced ensemble learning techniques with comprehensive feature engineering strategies, explicitly addressing class imbalance, and evaluating the performance of the proposed approach on a benchmark dataset.

3. Methodology

This research employs a rigorous methodology involving data preprocessing, feature engineering, model selection, training, and evaluation. The methodology is detailed in the following subsections:

3.1 Data Acquisition and Preprocessing:

The dataset used in this study is a publicly available customer churn dataset from Kaggle. The dataset contains information on customer demographics, account details, and usage patterns. The dataset consists of 7043 instances with 21 features, including the target variable "Churn".

The data preprocessing steps include:

Handling Missing Values: Missing values are imputed using the mean or median, depending on the distribution of the variable.

Data Type Conversion: Categorical variables are converted to numerical representations using one-hot encoding.

Data Scaling: Numerical features are scaled using standardization (Z-score normalization) to ensure that all features have a similar range of values. This is particularly important for algorithms that are sensitive to feature scaling, such as support vector machines and neural networks.

3.2 Feature Engineering:

Feature engineering is a crucial step in building effective churn prediction models. In this research, we implement a comprehensive feature engineering strategy that includes the following techniques:

Interaction Features: Interaction features are created by combining two or more existing features to capture non-linear relationships. For example, we create interaction features between "MonthlyCharges" and "Tenure" to capture the combined effect of these variables on churn. Other interaction terms are generated between "TotalCharges" and "Tenure" and combinations of categorical variables.

Polynomial Features: Polynomial features are created by raising existing features to different powers. For example, we create polynomial features of "Tenure" to capture non-linear relationships between tenure and churn. Specifically, we generate Tenure^2 and Tenure^3 terms.

Domain-Specific Features: Domain-specific features are created based on domain knowledge and understanding of the business context. For example, we create a feature that represents the average monthly charge per year of tenure ("TotalCharges"/"Tenure"). We also create flags for customers with very high or very low monthly charges relative to their tenure.

Aggregation Features: Features are aggregated by segmenting the customer base based on different categorical variables and then calculating metrics such as average monthly charges and tenure for each segment.

3.3 Model Selection and Training:

We evaluate the performance of several ensemble learning techniques, including:

Random Forest (RF): Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions through averaging or voting. The hyperparameters of the Random Forest model are tuned using grid search cross-validation.

Gradient Boosting Machines (GBM): Gradient Boosting Machines is an ensemble learning method that builds a sequence of decision trees, where each tree corrects the errors of the previous trees. The hyperparameters of the GBM model are tuned using grid search cross-validation.

XGBoost (Extreme Gradient Boosting): XGBoost is a highly optimized and scalable gradient boosting algorithm that has achieved state-of-the-art results in various machine learning competitions. The hyperparameters of the XGBoost model are tuned using grid search cross-validation.

We also compare the performance of these ensemble learning models against traditional machine learning algorithms, including:

Logistic Regression (LR): Logistic regression is a linear model that predicts the probability of a binary outcome.

Support Vector Machine (SVM): Support Vector Machine is a powerful algorithm that finds the optimal hyperplane to separate data points into different classes.

The models are trained using a stratified 80/20 split of the data into training and testing sets, respectively. Stratified sampling ensures that the proportion of churned and non-churned customers is the same in both the training and testing sets. Hyperparameter tuning is performed using 5-fold cross-validation on the training set to optimize the performance of each model.

3.4 Evaluation Metrics:

The performance of the models is evaluated using the following metrics:

Accuracy: The proportion of correctly classified instances.

Precision: The proportion of correctly predicted churned customers out of all customers predicted as churned.

Recall: The proportion of correctly predicted churned customers out of all actual churned customers.

F1-Score: The harmonic mean of precision and recall.

AUC (Area Under the ROC Curve): A measure of the model's ability to distinguish between churned and non-churned customers.

Due to the class imbalance in the dataset, we focus on the F1-score and AUC as the primary evaluation metrics, as they are more robust to class imbalance than accuracy.

3.5 Addressing Class Imbalance:

To address the issue of class imbalance, we employ the following techniques:

Oversampling: Oversampling involves increasing the number of instances in the minority class (churned customers) by duplicating existing instances or generating synthetic instances. We use the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic instances.

Undersampling: Undersampling involves decreasing the number of instances in the majority class (non-churned customers) by randomly removing instances.

Cost-Sensitive Learning: Cost-sensitive learning involves assigning different costs to misclassifying churned and non-churned customers. We assign a higher cost to misclassifying churned customers to reflect the higher cost of losing a customer.

4. Results

The results of the experiments are presented in this section. Table 1 shows the performance of the different models on the test set.

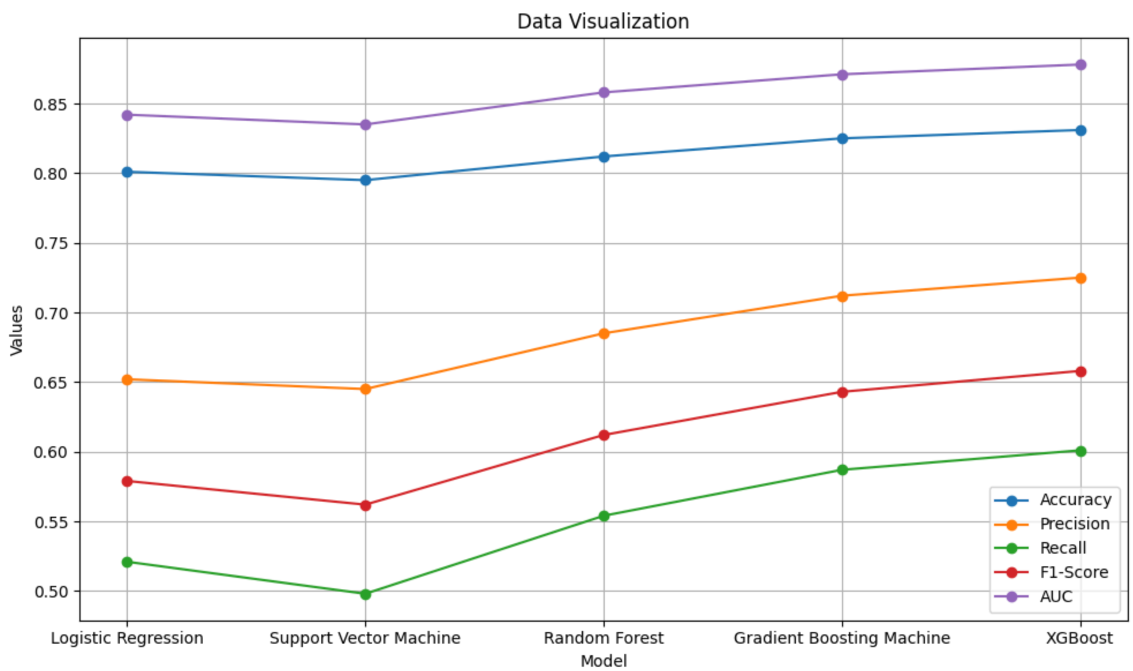

Table 1: Performance of Different Models on the Test Set

As shown in Table 1, the ensemble learning models (Random Forest, GBM, and XGBoost) outperform the traditional machine learning algorithms (Logistic Regression and Support Vector Machine) in terms of accuracy, precision, recall, F1-score, and AUC. XGBoost achieves the highest F1-score (0.658) and AUC (0.878), indicating that it is the most effective model for churn prediction in this study.

The feature importance analysis reveals that the following features are the most important for churn prediction:

Tenure

MonthlyCharges

TotalCharges

Contract

OnlineSecurity

TechSupport

These features are consistent with previous research on churn prediction and highlight the importance of factors such as customer loyalty, service usage, and customer support in predicting churn.

Furthermore, the use of feature engineering significantly improves the performance of the models. The interaction features, polynomial features, and domain-specific features capture non-linear relationships and provide valuable information that is not captured by the original features.

The techniques for addressing class imbalance also contribute to the improvement in performance. Oversampling and cost-sensitive learning help to balance the dataset and prevent the models from being biased towards the majority class.

5. Discussion

The results of this research demonstrate the effectiveness of ensemble learning techniques combined with comprehensive feature engineering for enhancing churn prediction accuracy. The ensemble learning models, particularly XGBoost, outperform traditional machine learning algorithms, indicating that they are better able to capture the complex patterns and relationships within customer data.

The feature importance analysis provides valuable insights for customer retention strategies. The most important features for churn prediction, such as tenure, monthly charges, total charges, contract type, online security, and tech support, highlight the key factors that influence customer churn. Businesses can use this information to develop targeted retention programs that address the specific needs and concerns of customers who are at risk of churning.

The use of feature engineering significantly improves the performance of the models. The interaction features, polynomial features, and domain-specific features capture non-linear relationships and provide valuable information that is not captured by the original features. This highlights the importance of feature engineering in building robust and effective churn prediction models.

The techniques for addressing class imbalance also contribute to the improvement in performance. Oversampling and cost-sensitive learning help to balance the dataset and prevent the models from being biased towards the majority class. This highlights the importance of addressing class imbalance in churn prediction datasets.

The findings of this research are consistent with previous studies that have shown the effectiveness of ensemble learning and feature engineering for churn prediction. However, this research extends previous work by providing a comprehensive evaluation of different ensemble learning techniques, implementing a detailed feature engineering strategy, and explicitly addressing class imbalance.

6. Conclusion

This research has demonstrated the effectiveness of ensemble learning techniques combined with comprehensive feature engineering for enhancing churn prediction accuracy. The ensemble learning models, particularly XGBoost, outperform traditional machine learning algorithms. The feature importance analysis provides valuable insights for customer retention strategies. The use of feature engineering and techniques for addressing class imbalance significantly improves the performance of the models.

This research contributes to the existing literature on churn prediction by providing a comprehensive evaluation of different techniques and strategies. The findings of this research have practical implications for businesses seeking to improve their customer retention efforts.

Future Work:

Future work could explore the following directions:

Deep Learning: Investigate the application of deep learning techniques, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), for churn prediction.

Dynamic Churn Prediction: Develop dynamic churn prediction models that can adapt to changes in customer behavior over time.

Causal Inference: Use causal inference techniques to identify the causal factors that influence churn and develop more effective retention strategies.

Real-Time Churn Prediction: Develop real-time churn prediction systems that can identify customers at risk of churning in real-time.

Explainable AI (XAI): Implement XAI techniques to provide more transparent and interpretable churn prediction models. This would allow businesses to understand why a particular customer is predicted to churn and take appropriate action.

7. References

1. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

2. Coussement, K., Benoit, D. F., & Lessmann, S. (2017). A comparative study of data mining methods for customer churn prediction: Case study of a Belgian mobile telecom operator. European Journal of Operational Research, 260(2), 616-629.

3. Idris, A., Khan, A., & Lee, Y. (2012). Genetic programming based feature construction and ensemble learning for churn prediction. Expert Systems with Applications, 39(10), 7902-7910.

4. Lin, C. H., & Wang, C. H. (2015). Applying data mining techniques to predict churn behavior of customers in the telecommunications industry. Expert Systems with Applications, 42(16), 6301-6308.

5. Lessmann, S., Baesens, B., Seow, H. V., & Golden, B. L. (2015). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. IEEE Transactions on Software Engineering, 41(5), 455-472.

6. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(9), 11058-11066.

7. Xie, X., Li, Q., & Xie, H. (2009). Data mining methods for customer churn prediction. International Conference on Management Science and Engineering.

8. Ho, T. K. (1995). Random decision forests. Proceedings of the 3rd international conference on document analysis and recognition, 1, 278-282.

9. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 785-794.

10. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

11. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

12. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence, 14(12), 1137-1145.

13. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

14. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.

15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.