DEPARTMENT OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING CONCORDIA UNIVERSITY

Predicting Customer Churn in a Subscription-Based Business

SOEN 6111 – Big Data Analytics

Winter 2025

Professor: Dr. Reza Mirsalari

Submitted by:

|  |  |
| --- | --- |
| Omer Sayem | 40226505 |
| Raymand Shojaie Aghabalaghe | 40258477 |
| Sepehr Zaki | 40280153 |

Date:

12 Feb 2025

Table of Contents

[Abstract 2](#_Toc189565030)

[Section 1: Introduction 2](#_Toc189565031)

[Section 2: Background and Literature Review 2](#_Toc189565032)

[Section 3: Proposed Model/ Framework 3](#_Toc189565033)

[Section 4: Contribution and Discussion 3](#_Toc189565034)

[Section 5: Conclusion 3](#_Toc189565035)

## **Abstract**

Customer churn prediction is a valuable business task that predicts customer retention and identifies customers’ pain points. It assists businesses in implementing critical strategies. In this report, we will explore the application of Decision Trees and Random Forests to predict customer churn for Stream Flex, a streaming service provider. By analyzing the customer demographics and distribution of subscription lengths, we built a predictive model that aids in business decision-making. We found that the Random Forest model outperforms Decision Trees in terms of accuracy (62% and 58%, respectively) and reliability, which makes it slightly preferred choice. The study also provides actionable business insights and recommendations based on key contributing factors identified in the model.

## **Section 1: Introduction**

**Problem Statement**

Many subscription-based businesses face the challenge of customer churn, where users cancel their subscriptions, resulting to revenue loss. If we can predict churn, companies can take proactive measures to retain customers. In our case, StreamFlex, a subscription-based streaming service, wants to identify churn-prone customers using machine learning techniques.

**Motivation**

Customer retention is more cost-effective than customer acquisition. We want to leverage predictive analytics; by which businesses can identify early warning signs of churn. Decision Trees and Random Forests, offer valuable insights into customer behavior and helps in taking strategic decision-making.

**Objectives**

* Analyze and preprocess customer churn data.
* Implement and compare Decision Tree and Random Forest models.
* Evaluate model performance using classification metrics.
* Derive business insights and suggest actionable strategies to minimize churn.

## **Section 2: Background and Literature Review (Application of Decision Trees in Business)**

* Why are decision trees useful in customer churn prediction?
  + While making critical business decisions, managers want to know many questions from the data. For example, has the customer logged in within the last 15 days? Did the customer face billing issues? From this closed binary question, businesses can make decisions. This is exactly how decision trees learn patterns from the data and can predict customer churn.
  + Decision trees split the data set based on the most important feature by leveraging concepts such as entropy (measure of disorder), information gain (how much uncertainty is reduced by the split), and regularization (pruning) to generalize new data.
* What business actions can be taken based on the predictions of a decision tree model?
  + Taking the result of the decision tree model, businesses can plan personalized promotions, discounts, and loyalty programs to the highly risky customers.
  + From the highest cause of churn, for example (logging issue), businesses can prioritize the task and give personalized solutions to the targeted customer.
  + Improve customer support and user experience and pivot the subscription plan.

## **Section 3: Proposed Model/ Framework**

* Data Preparation and Exploration
  + Load the given customer churn dataset.
  + Perform exploratory data analysis (EDA), including:
    - Summary Statistics (Mean, std, min, max, etc.)A screenshot of a computer

      AI-generated content may be incorrect.
    - Missing value handling by using
      * Median imputation for numeric values.
      * Mode imputation for categorical values.
    - Feature engineering:
      * Encoding the categorical values.
      * We have created new features from the existing data to capture hidden patterns from the data, and it enhances Model Interpretability:
        + **Support Interactions:** This is the summation of ). Higher values indicate poor customer experience.
        + **Engagement Score:** This is created by to analyze how much content a user consumes per login.
* Data visualization (Histograms, Box plots, Correlation matrix):
  + Distribution of Subscription Length:

A graph of a distribution of subscription length

AI-generated content may be incorrect.

* + Feature correlation matrix: A graph with red squares and blue squares

    AI-generated content may be incorrect.
* Decision Tree Classifier
  + Split the data into training and test sets (80% training, 20% testing)
  + Train a decision tree model using scikit-learn.
    - Feature scaling using StandardScaler: All data are scaled by **zero mean** and **unit variance** using the formula . Scaling (normalization) improves convergence speed.
    - Handing class imbalance using SMOTE: It generates synthetic data points for the minority class and helps to **prevent bias**.
  + Optimize hyperparameters using GridSearchCV:
    - Defining hyperparameter:
      * **max\_depth**: Limit the depth of the tree to prevent overfitting.
      * **min\_samples\_split:** Minimum number of samples required to split a node.
      * **min\_samples\_leaf:** Minimum number of samples of leaf.
    - GridSearchCV automatically finds the best combination of hyperparameters by training multiple model
    - Cross Validation: During training, the dataset is split into 5 sets, 4 for training 5th one for validating.
    - Scoring metrics: The Model minimizes **False positives**, optimizing the precision
  + Visualize the trained decision tree: The trained decision tree model is visualized using matplotlib.
  + Evaluate performance:
    - Accuracy: 58% (The model correctly classified 58% of the total prediction.)
    - Precision: 25.49% (From the class predicted churn, only 25.49% is actual churn)
    - Recall: 22.03% (The model detected only 22.03% of all actual churn)
    - F1-Score: 0.2364 (The balance between precision and recall)
    - Confusion matrix:
      * True Negative (TN) = 103
      * False Positive (FP) = 38
      * False Negative (FN) = 46
      * True Positive (TP) = 13

**The model has low precision and low recall; many of the classified churn customers are not churners.**

* Random Forest Classifier
  + Train a Random Forest model and compare its performance with Decision Trees.
    - The random forest model is an ensemble model that combines many decision trees to make classifications. It reduces overfitting and balance bias – variance.
    - Optimize hyperparameters using GridSearchCV.
    - Defining hyperparameter:
      * **n\_estimators**: Number of trees in the forest.
      * **max\_depth:** The maximum depth of each tree.
  + Analyze feature importance and determine key contributors to churn:

A graph with blue and white bars

AI-generated content may be incorrect.

**Customers who are unhappy and make complaints are highly likely to churn.**

* + Evaluate performance:
    - Accuracy: 62.5% (The model correctly classified 62.5% of the total prediction.)
    - Precision: 27.77% (From the class predicted churn, only 27.77% is actual churn)
    - Recall: 16.94% (The model detected only 16.94% of all actual churn)
    - F1-Score: 0.2105 (The balance between precision and recall)
    - Confusion matrix:
      * True Negative (TN) = 115
      * False Positive (FP) = 26
      * False Negative (FN) = 49
      * True Positive (TP) = 10
* Performance Evaluation
  + Compare Decision Tree and Random Forest models using classification metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 score |
| Decision tree | 0.58 | 0.254 | 0.220 | 0.2363 |
| Random forest | 0.625 | 0.277 | 0.1694 | 0.2105 |

* **Accuracy**: The Random Forest model outperformed the Decision Tree with a 62.5% accuracy.
* **Precision**: Slight improvement in detecting false positives in the Decision Tree model.
* **Recall**: Surprisingly, the Random Forest model’s performance on detecting all actual churners is reduced.
* **F1 Score**: As expected, the balance between precision and recall has reduced.

## **Section 4: Contribution and Discussion**

* Business Insights and Recommendations
  + Based on the model’s predictions, what characteristics contribute the most to customer churn?
    - **Number of Complaints**: Below standard customer service or delay in solving issue may drive customer to cancel their subscription.
    - **Subscription Length**: Customers who are on short subscriptions are likely to churn.
    - **Support Interactions (Complaints + Payment Issues + Resolution time):** Frequent complaints and longer resolution times strongly correlate with churn**.**
* What actionable insights can StreamFlex use to **reduce customer churn**?
  + Users with a high number of complaints and long resolution times are more likely to churn.
  + Users with **short subscriptions** are at higher risk of churn.
  + Users with short **watch times** and **infrequent logins** are at higher risk of churn.
  + Premium Users have a **lower churn rate** than basic members.
  + Payment issues contribute to churn, even for engaged users.
  + Users with **low engagement scores and multiple complaints** are at the highest risk.
* Suggest **three concrete business strategies** based on your findings.
  + Offer incentives to customers with high churn risk.
  + Improve customer support for those with frequent complaints.
  + Enhance subscription plans based on user preferences.

## **Section 5: Conclusion**

This report shines a light on the role of machine learning techniques (Decision tree, Random Forest) in customer churn prediction. By analyzing the results, we can recapitulate that the Random Forest model outperforms decision trees in terms of accuracy. However, a more powerful model like Gradient boosting (XGBoost, LightGBM) can be incorporated to get better accuracy. By identifying key churn factors, businesses can implement targeted retention strategies.

References

1. Zhang, B., 2023. Customer Churn in Subscription Business Model—Predictive Analytics on Customer Churn. *BCP Business & Management*, *44*, pp.870-876