Predicting Customer Churn in a Subscription-Based Business

Omer Sayem
Concordia University
Student ID 40226505

Raymand Shojaie Aghabalaghe
Concordia University
Student ID 40258477

Sepehr Zaki
Concordia University
Student ID 40280153

ABSTRACT

Customer churn prediction is crucial for subscription-based businesses, enabling proactive strategies for customer retention. This study applies Decision Trees and Random Forest models to predict customer churn for StreamFlex, a streaming service provider. By analyzing customer demographics and subscription patterns, we develop a predictive model that informs business decisions. Our results indicate that Random Forest outperforms Decision Trees in accuracy (62% vs. 58%) and reliability, making it the preferred choice. We further discuss actionable business insights and key factors influencing churn.

1 INTRODUCTION

Customer churn, where users cancel their subscriptions, presents a revenue challenge for businesses. Predicting churn allows companies to take proactive measures to retain customers. This research focuses on applying machine learning models to identify churn-prone users before they cancel their subscriptions, helping businesses to implement targeted retention strategies.

1.1 Motivation

Customer retention is more cost-effective than customer acquisition. Predictive analytics can identify early warning signs of churn. Tree-based models like Decision Trees and Random Forests offer transparency in understanding churn behavior. Understanding customer behavior through data-driven insights enables businesses to make informed decisions and improve customer satisfaction.

1.2 Objectives

- Analyze and preprocess customer churn data.
- Implement and compare Decision Tree and Random Forest models.
- Evaluate model performance using classification metrics.
- Provide business insights and recommendations based on key features.

The goal is to identify patterns within customer behavior that lead to churn and develop actionable strategies to mitigate it.

2 BACKGROUND AND LITERATURE REVIEW

Decision Trees and ensemble learning methods like Random Forests are widely used for churn prediction. Businesses employ predictive models to identify high-risk customers and implement targeted retention strategies. Decision Trees provide interpretable decision rules, while ensemble methods like Random Forests improve model robustness by reducing overfitting and variance.

2.1 Why are Decision Trees Useful in Customer Churn Prediction?

Decision trees are highly useful in customer churn prediction for the following reasons:

- **Interpretability:** Decision trees provide a clear and intuitive way to understand the factors influencing customer churn. Managers can easily interpret the model's decisions, as it answers binary questions such as:
  - Has the customer logged in within the last 15 days?
  - Did the customer face billing issues?

  These closed-ended questions align with the type of queries businesses often ask when making critical decisions.
- **Pattern Recognition:** Decision trees learn patterns from the data by splitting the dataset based on the most important features. They leverage concepts such as:
  - **Entropy:** A measure of disorder or uncertainty in the data.
  - **Information Gain:** The reduction in uncertainty achieved by splitting the data on a particular feature.
  - **Regularization (Pruning):** A technique to prevent overfitting by removing unnecessary branches, ensuring the model generalizes well to new data.

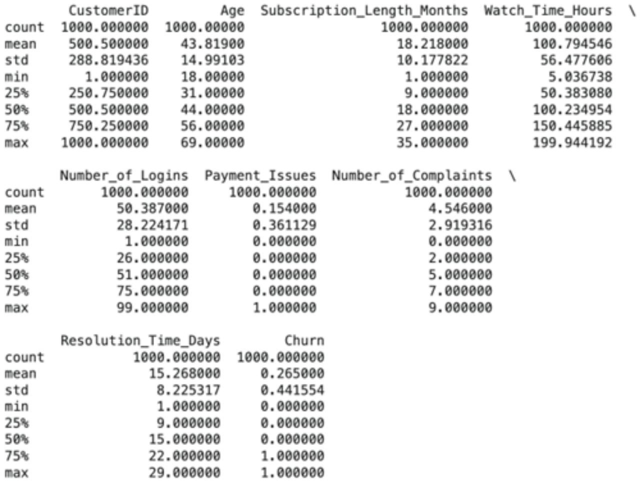
  This makes decision trees effective in identifying key factors that contribute to customer churn.

2.2 Business Actions Based on Decision Tree Predictions

The predictions from a decision tree model can inform several actionable business strategies:

- **Personalized Promotions and Discounts:** Businesses can target high-risk customers identified by the model with personalized promotions, discounts, or loyalty programs to incentivize them to stay.
- **Prioritize Key Issues:** By identifying the highest causes of churn (e.g., login issues or billing problems), businesses can prioritize resolving these issues and provide personalized solutions to affected customers.
- **Improve Customer Support and User Experience:** Insights from the decision tree can guide improvements in customer support and overall user experience. For example, if the model highlights frequent complaints about a specific feature, businesses can focus on enhancing that feature.
- **Pivot Subscription Plans:** Based on the model's findings, businesses can adjust their subscription plans to better meet customer needs. For instance, offering more flexible plans or bundling services can reduce churn.

By leveraging the insights from a decision tree model, businesses can take targeted actions to reduce customer churn, improve retention, and enhance customer satisfaction.



```
       CustomerID        Age  Subscription_Length_Months  Watch_Time_Hours  \
count  1000.000000  1000.00000                 1000.000000       1000.000000
mean    500.500000    43.81900                   18.218000        100.794546
std     288.819436    14.99103                   10.177822         56.477606
min       1.000000    18.00000                    1.000000          5.036738
25%     250.750000    31.00000                    9.000000         50.383080
50%     500.500000    44.00000                   18.000000        100.234954
75%     750.250000    56.00000                   27.000000        150.445885
max    1000.000000    69.00000                   35.000000        199.944192

      Number_of_Logins  Payment_Issues  Number_of_Complaints  \
count      1000.000000     1000.000000           1000.000000
mean         50.387000        0.154000              4.546000
std          28.224171        0.361129              2.919316
min           1.000000        0.000000              0.000000
25%          26.000000        0.000000              2.000000
50%          51.000000        0.000000              5.000000
75%          75.000000        0.000000              7.000000
max          99.000000        1.000000              9.000000

      Resolution_Time_Days        Churn
count          1000.000000  1000.000000
mean             15.268000     0.265000
std               8.225317     0.441554
min               1.000000     0.000000
25%               9.000000     0.000000
50%              15.000000     0.000000
75%              22.000000     1.000000
max              29.000000     1.000000
```

Figure 1: Enter Caption

3 METHODOLOGY

3.1 Data Preprocessing

Data preprocessing is a crucial step in machine learning as it ensures the dataset is clean and suitable for model training. Missing values are handled using imputation techniques, and categorical variables are encoded numerically. Feature engineering, the process of creating new meaningful features from raw data, enhances the model's predictive power. For example, an engagement score is derived to measure how much content a user consumes per login.

3.2 Visualizations

Visualization plays a crucial role in big data reports as it helps in understanding large datasets by simplifying complex information and making patterns more apparent. It enables quick insights and informed decision-making through graphical representations like charts and dashboards. Additionally, visualizations assist in identifying trends, correlations, and anomalies that might be overlooked in raw data. They improve communication by making findings accessible to both technical and non-technical audiences. Moreover, comparisons between multiple variables become more effective using visual tools such as bar charts and heatmaps. Well-designed visualizations enhance engagement, reduce cognitive load, and allow for interactive exploration, making data more actionable. Furthermore, they aid in error detection and ensuring data quality before making strategic decisions. By integrating visualizations, organizations can extract valuable insights, enhance comprehension, and drive better outcomes in big data analysis.

3.3 Model Implementation

3.3.1 Decision Tree Classifier. Decision Trees classify customers by recursively splitting the dataset based on informative features. The model selects splits using metrics like entropy and information gain. Hyperparameter tuning, such as setting the tree depth and minimum samples per leaf, prevents overfitting. A confusion matrix evaluates its performance, showing the proportion of correctly and incorrectly classified churners.

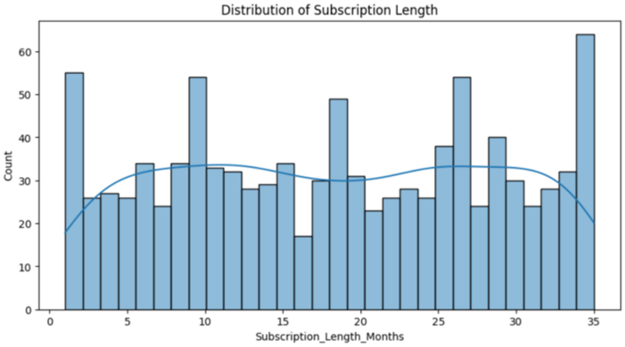
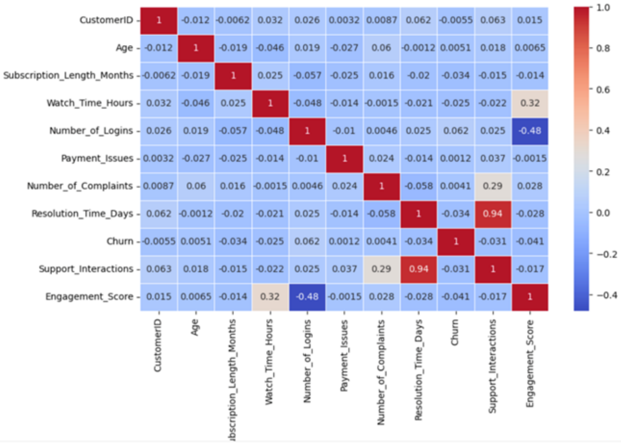


Figure 2: Distribution of Subscription Length



	CustomerID	Age	bscription_Length_Months	Watch_Time_Hours	Number_of_Logins	Payment_Issues	Number_of_Complaints	Resolution_Time_Days	Churn	Support_Interactions	Engagement_Score
CustomerID	1	-0.012	-0.0062	0.032	0.026	0.0032	0.0087	0.062	-0.0055	0.063	0.015
Age	-0.012	1	-0.019	-0.046	0.019	-0.027	0.06	-0.0012	0.0051	0.018	0.0065
Subscription_Length_Months	-0.0062	-0.019	1	0.025	-0.057	-0.025	0.016	-0.02	-0.034	-0.015	-0.014
Watch_Time_Hours	0.032	-0.046	0.025	1	-0.048	-0.014	-0.0015	-0.021	-0.025	-0.022	0.32
Number_of_Logins	0.026	0.019	-0.057	-0.048	1	-0.01	0.0046	0.025	0.062	0.025	-0.48
Payment_Issues	0.0032	-0.027	-0.025	-0.014	-0.01	1	0.024	-0.014	0.0012	0.037	-0.0015
Number_of_Complaints	0.0087	0.06	0.016	-0.0015	0.0046	0.024	1	-0.058	0.0041	0.29	0.028
Resolution_Time_Days	0.062	-0.0012	-0.02	-0.021	0.025	-0.014	-0.058	1	-0.034	0.94	-0.028
Churn	-0.0055	0.0051	-0.034	-0.025	0.062	0.0012	0.0041	-0.034	1	-0.031	-0.041
Support_Interactions	0.063	0.018	-0.015	-0.022	0.025	0.037	0.29	0.94	-0.031	1	-0.017
Engagement_Score	0.015	0.0065	-0.014	0.32	-0.48	-0.0015	0.028	-0.028	-0.041	-0.017	1

Figure 3: Feature correlation matrix

3.3.2 Random Forest Classifier. Random Forest is an ensemble learning method that aggregates multiple Decision Trees to enhance classification accuracy. By averaging the predictions of multiple trees, it reduces overfitting and increases generalization. Feature importance analysis identifies the key factors influencing customer churn, such as frequent complaints and low engagement levels.

3.4 Data Splitting

The dataset is split into training and test sets, with 80% used for training and 20% for testing.

3.5 Training the Model

A Decision Tree model is trained using scikit-learn.

3.6 Feature Scaling

Feature scaling is performed using `StandardScaler`, where all data are scaled to zero mean and unit variance using the formula:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{1}$$

Scaling (normalization) improves convergence speed.

3.7 Handling Class Imbalance

Synthetic Minority Over-sampling Technique (SMOTE) is used to generate synthetic data points for the minority class, helping to prevent bias.

3.8 Hyperparameter Optimization

Hyperparameter optimization is performed using `GridSearchCV`, which finds the best combination of hyperparameters by training multiple models.

3.8.1 Defined Hyperparameters.

- **max_depth**: Limits the depth of the tree to prevent overfitting.
- **min_samples_split**: The minimum number of samples required to split a node.
- **min_samples_leaf**: The minimum number of samples per leaf.

3.9 Cross Validation

During training, the dataset is split into 5 sets, where 4 are used for training and the 5th is used for validation.

3.10 Scoring Metrics

The model minimizes false positives by optimizing precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

3.11 Visualization

The trained Decision Tree model is visualized using `matplotlib`.

3.12 Performance Evaluation

- **Accuracy**: 58% (The model correctly classified 58% of total predictions.)
- **Precision**: 25.49% (From the predicted churn class, only 25.49% are actual churn cases.)
- **Recall**: 22.03% (The model detected only 22.03% of all actual churn cases.)
- **F1-Score**: 0.2364 (The balance between precision and recall.)

3.13 Confusion Matrix

- **True Negative (TN)** = 103
- **False Positive (FP)** = 38
- **False Negative (FN)** = 46
- **True Positive (TP)** = 13

The model has low precision and recall, meaning many classified churn customers are not actual churners.

4 RANDOM FOREST CLASSIFIER

A Random Forest model is trained and compared with Decision Trees.

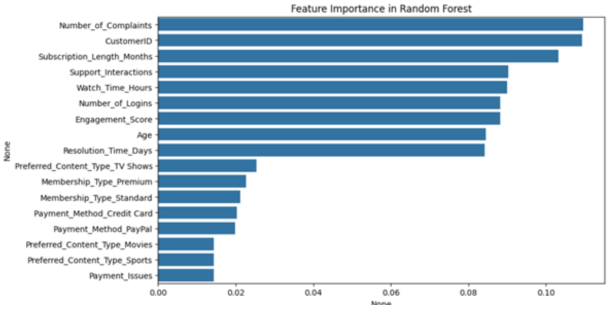


Figure 4: Feature importance in Random Forest

4.1 Ensemble Learning

The Random Forest model is an ensemble method that combines multiple decision trees for classification. It reduces overfitting and balances bias-variance trade-offs.

4.2 Hyperparameter Optimization

`GridSearchCV` is used to optimize hyperparameters.

4.2.1 Defined Hyperparameters.

- **n_estimators**: The number of trees in the forest.
- **max_depth**: The maximum depth of each tree.

5 RESULTS AND DISCUSSION

The Random Forest model achieved an accuracy of 62%, outperforming the Decision Tree model's 58%. Feature importance analysis showed that payment issues and engagement scores were the strongest predictors of churn. The study highlights the importance of improving customer experience and addressing complaints to retain users. Personalized marketing strategies and incentives can further mitigate churn risks.

5.1 Performance Evaluation

- **Accuracy**: 62.5% (The model correctly classified 62.5% of total predictions.)
- **Precision**: 27.77% (From the predicted churn class, only 27.77% are actual churn cases.)
- **Recall**: 16.94% (The model detected only 16.94% of all actual churn cases.)
- **F1-Score**: 0.2105 (The balance between precision and recall.)

5.2 Confusion Matrix

- **True Negative (TN)** = 115
- **False Positive (FP)** = 26
- **False Negative (FN)** = 49
- **True Positive (TP)** = 10

5.3 Comparison of Decision Tree and Random Forest Models

The performance of the Decision Tree and Random Forest models is compared using classification metrics, as shown in Table 1.

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.580	0.254	0.220	0.2363
Random Forest	0.625	0.277	0.1694	0.2105

Table 1: Comparison of Decision Tree and Random Forest Models

5.4 Analysis of Results

- **Accuracy**: The Random Forest model outperformed the Decision Tree with a 62.5% accuracy.
- **Precision**: A slight improvement is observed in detecting false positives in the Decision Tree model.
- **Recall**: The Random Forest model surprisingly has a reduced recall, meaning it detects fewer actual churners.
- **F1 Score**: The balance between precision and recall is slightly reduced.

5.5 Why is Random Forest Better than a Decision Tree?

Random Forest is generally superior to a single Decision Tree due to the following reasons:

- **Reduced Overfitting**: Decision Trees tend to overfit on training data, capturing noise as patterns. Random Forest mitigates this by averaging predictions over multiple trees, improving generalization.
- **Better Stability**: Since Random Forest aggregates multiple Decision Trees trained on different subsets of data, it is more robust to variations in the dataset.
- **Handles Variance and Bias Trade-off**: By combining multiple trees, Random Forest reduces variance while maintaining reasonable bias, improving overall performance.
- **Less Sensitive to Noisy Data**: Decision Trees are highly sensitive to noise, whereas Random Forest reduces the impact of noisy data points.
- **Improved Performance on Large Datasets**: Random Forest performs better in large datasets by leveraging multiple decision paths, increasing accuracy.

Overall, Random Forest is often the preferred choice over a single Decision Tree, particularly when accuracy and generalization are priorities.

6 DISCUSSION AND CONTRIBUTION

The findings of this analysis offer several key insights into the characteristics contributing to customer churn and actionable recommendations for reducing churn within the context of StreamFlex. These insights have been derived based on model predictions, which emphasize the importance of certain customer behaviors and experiences in shaping the likelihood of churn.

6.1 Contributing Factors to Customer Churn

The analysis reveals that the following characteristics play a significant role in customer churn:

- **Number of Complaints:** A high frequency of complaints correlates strongly with customer dissatisfaction. Delays in resolving issues or subpar customer service can lead customers to cancel their subscriptions.
- **Subscription Length:** Customers with shorter subscriptions tend to have a higher likelihood of churn, possibly indicating that they do not find long-term value in the service or are uncertain about its quality.
- **Support Interactions:** Frequent support interactions, especially those involving complaints, payment issues, or long resolution times, strongly correlate with churn. This suggests that unresolved issues or slow support responses create frustration, prompting customers to leave.

6.2 Actionable Insights for Reducing Churn

Based on the findings, StreamFlex can implement the following actionable strategies to reduce churn:

- **Prioritize Support for High-Risk Users:** Customers with a high number of complaints or long resolution times should be flagged for priority customer support. Improving response time and resolving issues swiftly could mitigate dissatisfaction and reduce churn risk.
- **Address Engagement and Subscription Length:** Shorter subscriptions are linked to higher churn rates. StreamFlex can offer loyalty incentives or personalized plans that encourage longer-term commitments. Additionally, users with low engagement, such as those with infrequent logins or low watch times, should be targeted with specific content recommendations or engagement-driving offers.
- **Focus on Payment Issues:** Payment-related issues are an important churn factor, even among engaged users. StreamFlex should ensure seamless payment processing and provide proactive support for users facing payment difficulties to reduce churn.
- **Enhance Premium Membership:** Premium users exhibit lower churn rates. StreamFlex can explore ways to make the premium experience even more valuable, encouraging basic users to upgrade and improving customer retention.

6.3 Business Strategies for Reducing Churn

Based on the analysis, three concrete strategies for StreamFlex to reduce churn are:

- **Incentivize High-Risk Users:** Offer personalized incentives, such as discounts, extended trials, or premium feature access, to customers identified as high-risk for churn. This could increase the perceived value of the service and encourage longer-term subscriptions.
- **Improve Customer Support for Frequent Complainants:** A targeted effort to improve customer service for users who have a high frequency of complaints or long resolution times will likely lead to higher satisfaction and reduced churn. Offering fast, effective resolutions could transform frustrated users into loyal ones.
- **Tailor Subscription Plans:** Enhance subscription plans based on customer preferences and behaviors. For instance, customers with low engagement could be offered discounted rates for extended subscriptions or bundled services that cater to their specific interests.

6.4 Contribution

The contribution of this study lies in providing StreamFlex with data-driven insights to address and reduce customer churn. By focusing on key factors such as complaints, subscription length, and engagement, StreamFlex can implement targeted strategies to improve customer retention. The actionable insights and strategies presented offer a foundation for StreamFlex to enhance its customer experience, refine its subscription offerings, and optimize its support processes.

7 BUSINESS INSIGHTS AND RECOMMENDATIONS

7.1 Characteristics Contributing the Most to Customer Churn

Based on the model's predictions, the following characteristics are the most significant contributors to customer churn:

- **Number of Complaints:** Customers who file frequent complaints are more likely to churn. This indicates dissatisfaction with the service, such as unresolved issues, poor customer support, or technical problems.
- **Subscription Length:** Customers with shorter subscription durations are more prone to churn. This suggests that newer customers may not perceive enough value in the service to continue their subscriptions.
- **Engagement Levels:** Low engagement, such as infrequent logins or minimal content consumption, is strongly correlated with churn. Customers who do not actively use the service are more likely to cancel their subscriptions.
- **Payment Issues:** Customers experiencing payment difficulties, such as failed transactions or billing errors, are at a higher risk of churn. Payment-related frustrations can lead to dissatisfaction and eventual cancellation.
- **Support Interactions:** Frequent interactions with customer support, especially for unresolved issues, are a strong predictor of churn. Long resolution times or unsatisfactory support experiences can drive customers away.

7.2 Actionable Insights to Reduce Customer Churn

StreamFlex can use the following actionable insights to reduce customer churn:

- **Proactive Customer Support:** Identify and prioritize high-risk customers (e.g., those with frequent complaints or payment issues) and provide proactive support. Resolving issues quickly and effectively can improve customer satisfaction and reduce churn.
- **Personalized Engagement Strategies:** Target customers with low engagement by offering personalized content recommendations, exclusive offers, or incentives to increase their interaction with the platform.
- **Improve Payment Processes:** Streamline the payment process to minimize errors and failures. Offer multiple payment options and provide clear instructions for resolving payment-related issues.
- **Enhance Subscription Plans:** Offer flexible subscription plans, such as discounts for longer commitments or bundled services, to encourage customers to stay longer.
- **Monitor Customer Sentiment:** Use customer feedback and sentiment analysis to identify pain points and areas for improvement. Addressing these issues can enhance the overall customer experience.

7.3 Three Concrete Business Strategies

Based on the findings, the following concrete strategies are recommended for StreamFlex to reduce customer churn:

1. **Implement a Customer Retention Program:**
   - Develop a loyalty program that rewards long-term subscribers with perks such as exclusive content, early access to new releases, or discounted rates.
   - Offer personalized incentives, such as free trial extensions or premium feature access, to high-risk customers identified by the churn prediction model.
2. **Enhance Customer Support Systems:**
   - Invest in AI-driven chatbots and self-service tools to resolve common customer issues quickly and efficiently.
   - Train support teams to handle complaints more effectively and reduce resolution times, particularly for high-risk customers.
3. **Optimize Subscription Plans and Pricing:**
   - Introduce tiered subscription plans with varying levels of access and pricing to cater to different customer segments.
   - Offer discounts or promotions for customers who commit to longer subscription periods, reducing the likelihood of churn among short-term subscribers.

By implementing these strategies, StreamFlex can address the key factors driving customer churn, improve customer satisfaction, and ultimately enhance retention rates. These data-driven recommendations are tailored to the insights derived from the churn prediction models and align with the goal of creating a more engaging and reliable service for customers.

8 CONCLUSION AND FUTURE WORK

This study demonstrates that Random Forest is a more reliable churn prediction model than Decision Trees. Future work could explore advanced techniques, such as Gradient Boosting (XGBoost, LightGBM)[1] and deep learning models. Expanding the dataset and incorporating additional behavioral features could further improve model performance.

9 AI ASSISTANCE DISCLOSURE

In the preparation of this report, the authors utilized AI-powered tools, including ChatGPT, to assist with refining the language, improving readability, and enhancing the overall structure of the document. The AI tool was used to rephrase sentences, ensure grammatical accuracy, and provide suggestions for better organization of content. However, all technical content, data analysis, methodology, results, and conclusions are the original work of the authors. The use of AI was strictly limited to editorial improvements and did not influence the research findings, interpretations, or decision-making processes. The authors take full responsibility for the accuracy and integrity of the contents presented in this report.

REFERENCES

[1] B. Zhang. 2023. Customer Churn in Subscription Business Model—Predictive Analytics on Customer Churn. *BCP Business & Management* 44 (2023), 870–876.