UNIVERSITÉ Concordia UNIVERSITY

GINA CODY SCHOOL OF ENGINEERING AND COMPUTER SCIENCE

SOEN 6111 W25

Assignment 1 Report

Group 54

This report is written as part of Assignment 1: Predicting Customer Churn in a Subscription-Based Business for Big Data Analytics course. The goal of this assignment is to examine customer churn patterns by training models, notably like Decision Trees and Random Forests. The report describes the process of exploratory data analysis, model construction, performance evaluation, and business insights gained from the study.

Group Members

Name	ID
Mihir Panchal	40291315
Yashesh sorathia	40267022
Abhi Pareshbhai Patel	40289176

Theoretical Understanding

Section 1: Application of Decision Trees in Business (Theoretical Analysis)

1. Why are Decision Trees useful in customer churn prediction?

Decision trees are useful for predicting customer turnover because of their interpretability, versatility, and ability to handle a variety of data types. Key benefits include:

• **Simple Interpretation:** They offer a straightforward, rule-based assessment of churn likelihood, making insights actionable.

• **Handling Mixed Data:** Capable of processing numerical and categorical characteristics such as watch time and payment methods.

• **Identifying Key Factors:** Identifies the most important churn indicators, such as payment troubles or low engagement.

• **Modelling Complex Patterns:** Creates non-linear correlations between customer attributes and churn probability.

• **Resilient to Missing Data:** Can split based on available qualities, resulting in robust predictions.

Decision Trees are an effective baseline model, providing immediate and interpretable insights regarding customer turnover behaviour.

2. What business actions can be taken based on Decision Tree predictions?

Businesses can utilize churn forecasts to develop targeted retention strategies.

• **Personalized Engagement:** Offer special material, discounts, or recommendations to at-risk clientele.

• **Payment Recovery:** Follow up on missed payments with reminders, flexible options, or one-time savings.

• **Proactive Support:** Prioritize high-risk consumers for faster issue resolution and specialist assistance.

• **Subscription Optimisation:** Provide plan upgrades or free trials to users who are likely to cancel their base subscription.

• **Re-engagement campaigns:** Use emails, reminders, and loyalty programs to retain disengaged customers.

Businesses that apply Decision Tree insights can reduce churn, improve customer satisfaction, and increase retention.

Section 2: Python Implementation – Building the Model

This section is divided further into total four tasks to build the model as following:

1. Data Exploration & Preprocessing

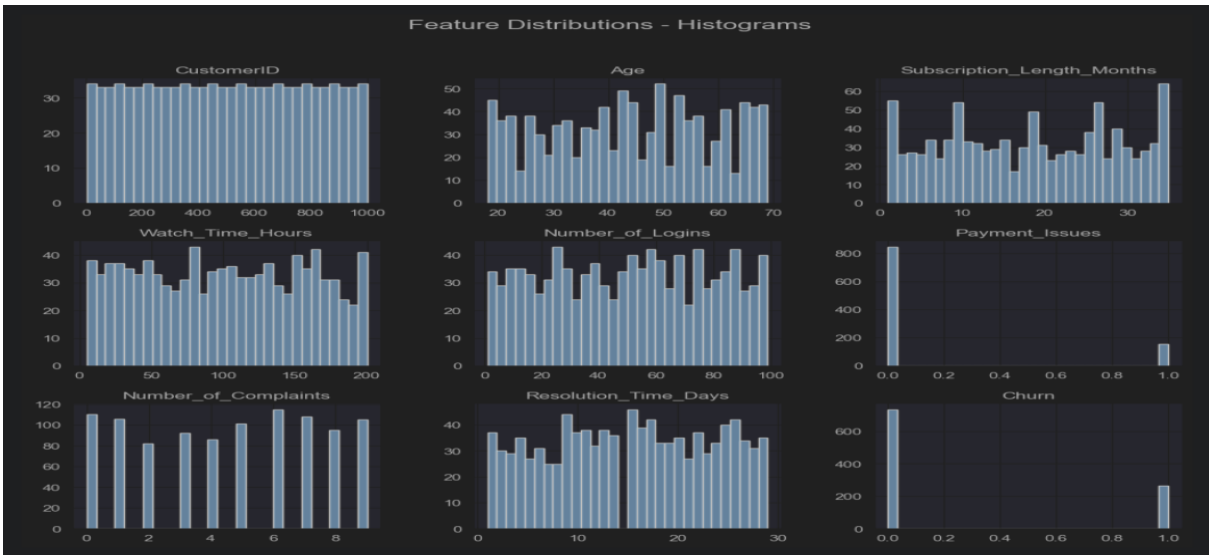
• The dataset was loaded using read_csv() function by importing Panda's library and analyzed for summary statistics, missing values, and data distributions using different functions like df.head(), display(df.info()) and display(df.isnull().sum()).

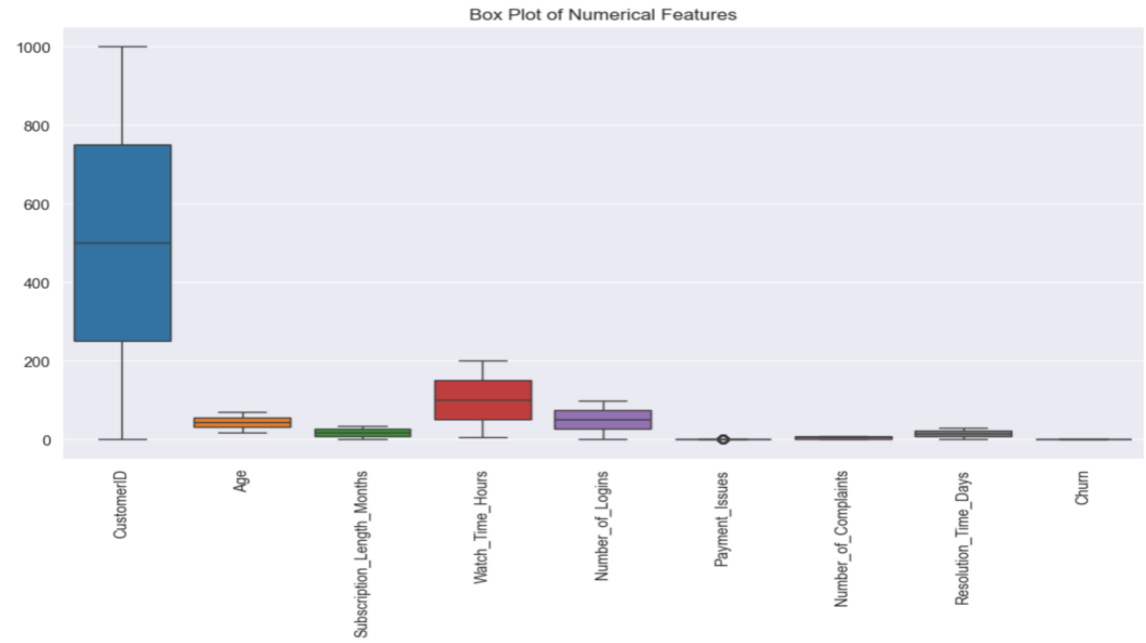
Summary Statistics:

	CustomerID	Age	Subscription_Length_Months	Watch_Time_Hours	Number_of_Logins	Payment_Issues	Number_of_Complaints
count	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000
mean	500.500000	43.81900	18.218000	100.794546	50.387000	0.154000	4.546000
std	288.819436	14.99103	10.177822	56.477606	28.224171	0.361129	2.919316
min	1.000000	18.00000	1.000000	5.036738	1.000000	0.000000	0.000000
25%	250.750000	31.00000	9.000000	50.383080	26.000000	0.000000	2.000000
50%	500.500000	44.00000	18.000000	100.234954	51.000000	0.000000	5.000000
75%	750.250000	56.00000	27.000000	150.445885	75.000000	0.000000	7.000000
max	1000.000000	69.00000	35.000000	199.944192	99.000000	1.000000	9.000000

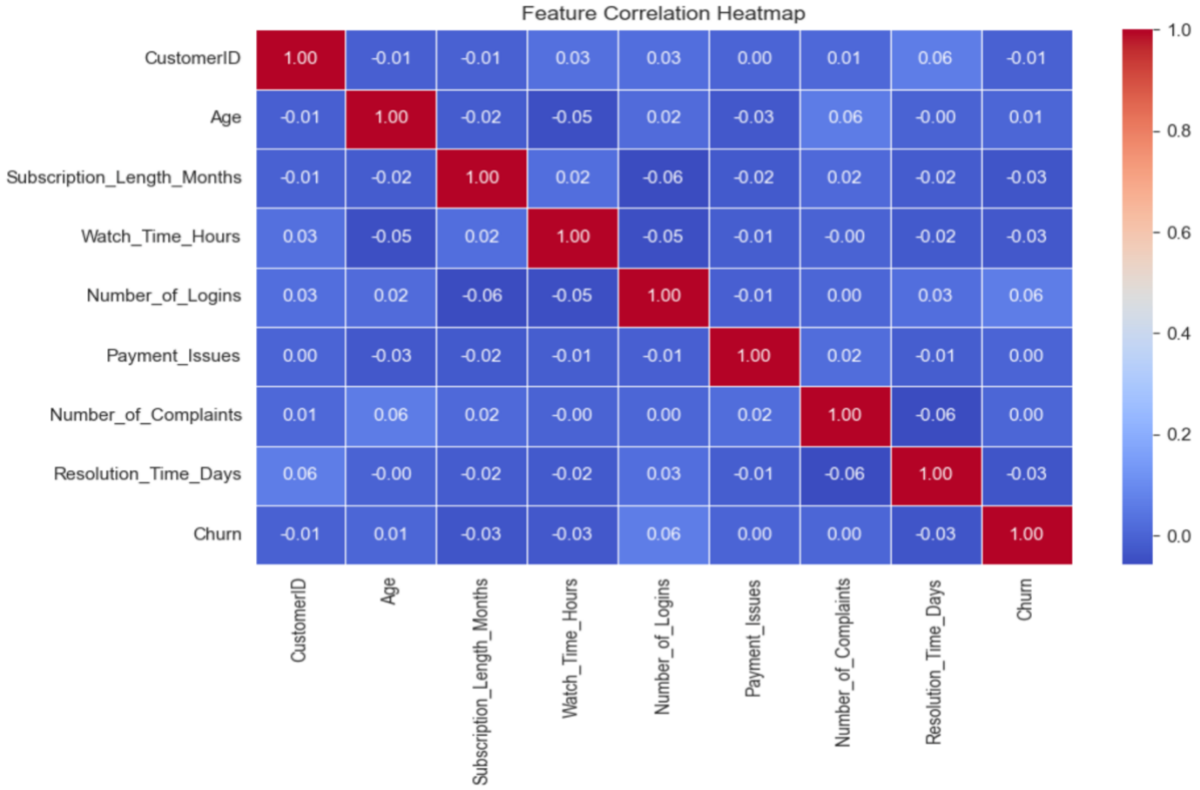
```
Missing Values:
CustomerID                  0
Age                         0
Subscription_Length_Months  0
Watch_Time_Hours            0
Number_of_Logins            0
Preferred_Content_Type      0
Membership_Type             0
Payment_Method              0
Payment_Issues              0
Number_of_Complaints        0
Resolution_Time_Days        0
Churn                       0
dtype: int64
<Figure size 1200x1000 with 0 Axes>
```

• Histograms as per below shows that watch time, number of logins, and subscription length show varied distributions, indicating different usage patterns among users. Payment issues and churn are highly imbalanced, with most users not experiencing payment issues but a significant number of users churning.





• Correlation heatmap shows no strong correlations between churn and other features, suggesting churn is influenced by multiple small factors. Subscription length and number of logins have a weak negative correlation with churn, indicating higher engagement reduces the likelihood of churn. Payment issues have little correlation with churn, suggesting that other factors might be more influential in customer retention.



2. Decision Tree Model Implementation

• First the dataset was split into training and testing sets, ensuring a balanced evaluation.

```
print(f"Training Set Size: {X_train.shape}, Test Set Size: {X_test.shape}")
```

```
Training Set Size: (800, 11), Test Set Size: (200, 11)
```

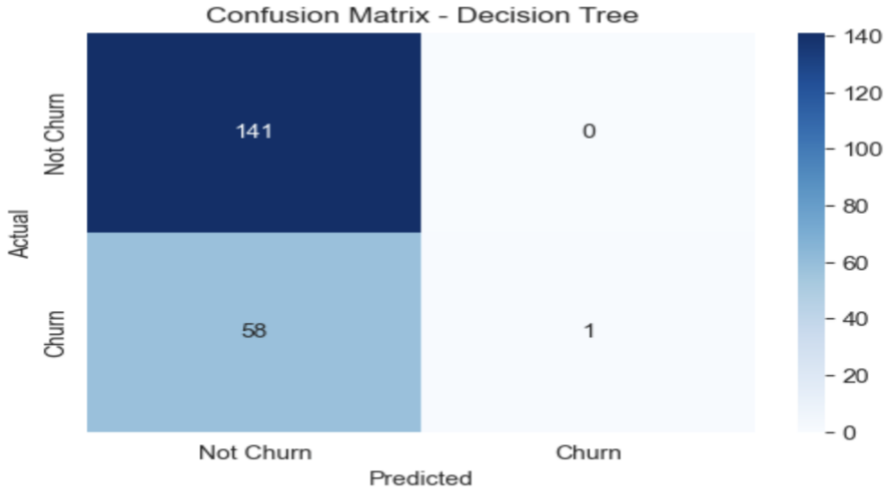
• Then a Decision Tree Classifier was trained using scikit-learn, with hyperparameter tuning via GridSearchCV to optimize performance.

```
# Use the best model for predictions
best_dt_classifier = grid_search.best_estimator_
y_pred_dt = best_dt_classifier.predict(X_test)
```

```
Best Parameters: {'criterion': 'entropy', 'max_depth': 3, 'min_samples_leaf': 4, 'min_samples_split': 2}
Best Cross-Validation Accuracy: 0.72625
```

• The model's effectiveness was measured using accuracy, precision, recall, F1-score, and a confusion matrix. The Decision Tree model achieves 71% accuracy. The confusion matrix reveals that it correctly labels all 141 "Not Churn" situations but incorrectly classifies 58 of 59 true churn cases. The precision for churn is 100%, but the recall is severely low at 2%, yielding a terrible F1-score of 0.03.

```
Decision Tree Performance:
Accuracy: 0.71
Precision: 1.0
Recall: 0.01694915254237288
F1 Score: 0.0333333333333333
```



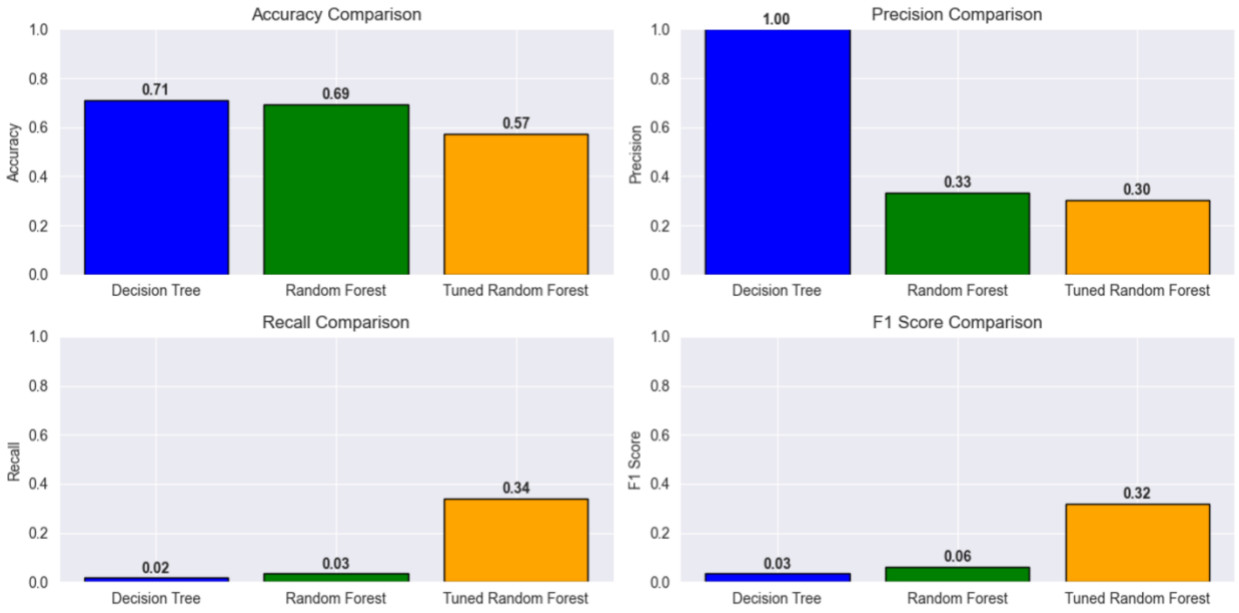

```
Classification Report:
              precision    recall  f1-score   support

   Not Churn       0.71      1.00      0.83       141
       Churn       1.00      0.02      0.03        59

    accuracy                           0.71       200
   macro avg       0.85      0.51      0.43       200
weighted avg       0.79      0.71      0.59       200
```

3. Random Forest Model for Performance Improvement

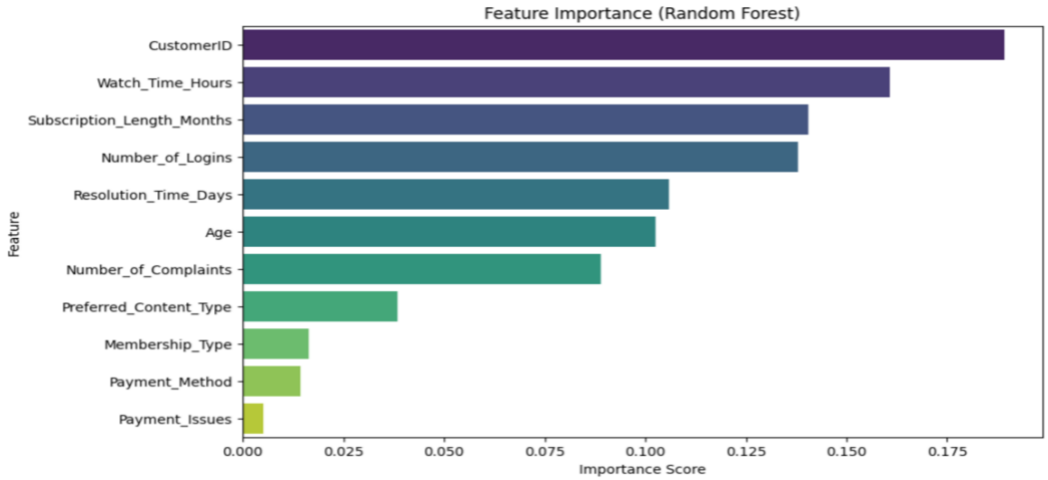
• A Random Forest Classifier was implemented to enhance prediction accuracy. In this we have implemented one more thing by ourselves is that we have added two types of random forest one that gives result before applying hyperparameter tuning and another after it.



Here, Decision Tree has the highest accuracy, but it's overfitting and fails on recall. Random Forest before tuning performs slightly worse than Decision Tree in accuracy but improves recall. Tuned Random Forest significantly improves recall (0.34), making it better at identifying churned customers. F1-score increases after hyperparameter tuning, balancing precision and recall.

4. Business Insights & Recommendations

• Below chart shows the customer behaviors and attributes contributing to churn.



According to above graph, the top factors impacting customer churn are watch time, subscription length, number of logins, and complaint resolution time.

To reduce churn, StreamFlex should implement the following strategies:

1. Increase **User Engagement** with personalized content and rewards. To improve retention, tailored recommendations based on viewing behaviours should be provided. Use AI-powered content suggestions to increase engagement. Encourage frequent platform use by offering streak incentives for consecutive logins.

2. Improve **Customer Support & Issue Resolution Time**. Slow complaint resolution is a big element that leads to user abandonment. Implement chatbots for quick troubleshooting and prioritize high-risk consumers for prompt human assistance.

3. Improve **Subscriber Retention & Payment Flexibility**. To encourage long-term subscriptions, give discounts for annual memberships. Provide various payment choices and reminders to reduce churn from payment failures.

Three concrete business strategies are as follows:

1. Use AI-driven recommendations and push notifications for Personalized Content & Notifications to encourage frequent platform usage.

2. Also add priority support & faster issue resolution by implementing a tiered support system to resolve complaints quickly.

3. Loyalty & retention programs also helps by offering discounts or perks for long-term subscriptions to encourage customer commitment.