SOEN 6111-Big Data Analytics

Assignment 1

Group -12

Tanzina Nasrin – 40235506
Tania Sanjid – 40255010
Md Abdul Hai - 40270829

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

Why are Decision Trees Useful in Customer Churn Prediction?

Decision trees are a popular machine learning algorithm for classification tasks, such as predicting customer churn, due to the following reasons:

1. **Interpretability**: Decision trees are easy to understand and interpret. They provide a clear visual representation of the decision-making process, which is crucial for business stakeholders who may not have a technical background.
2. **Handling Non-Linear Relationships**: Decision trees can capture non-linear relationships between features and the target variable, making them suitable for complex datasets.
3. **Feature Importance**: Decision trees can rank features based on their importance in predicting the target variable, helping businesses identify key factors contributing to churn.
4. **Handling Mixed Data Types**: Decision trees can handle both numerical and categorical data without requiring extensive preprocessing.
5. **Robustness to Outliers**: Decision trees are less sensitive to outliers compared to other algorithms like linear regression.

What Business Actions Can Be Taken Based on the Predictions of a Decision Tree Model?

Based on the predictions of a decision tree model, businesses can take the following actions:

1. **Targeted Retention Campaigns**: Identify customers who are likely to churn and offer them personalized incentives, such as discounts, free trials, or exclusive content, to retain them.
2. **Improve Customer Support**: If the model identifies that customers with a high number of complaints are more likely to churn, the business can improve its customer support system to resolve issues more efficiently.
3. **Product and Service Enhancements**: If certain features (e.g., preferred content type, payment issues) are identified as significant contributors to churn, the business can focus on improving those areas.
4. **Subscription Plan Adjustments**: If customers on certain subscription plans (e.g., Basic) are more likely to churn, the business can consider revising the pricing or features of those plans.
5. **Proactive Engagement**: Use the model to identify at-risk customers and engage with them proactively through personalized communication or offers.

Section 2: Python Implementation – Building the Model

Task 1: Data Preparation and Exploration

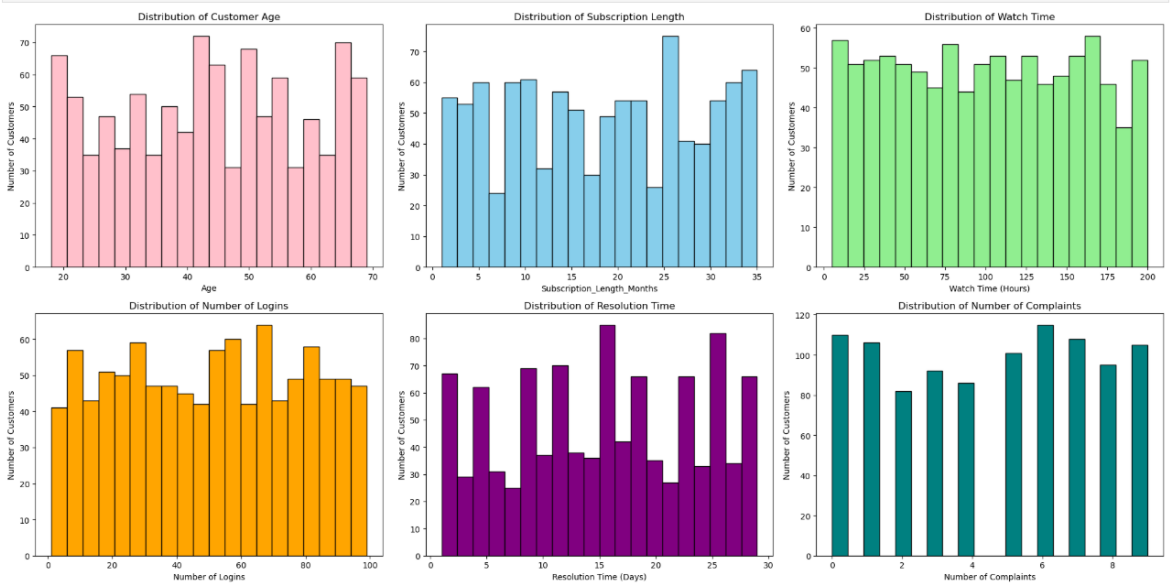
Load the Dataset

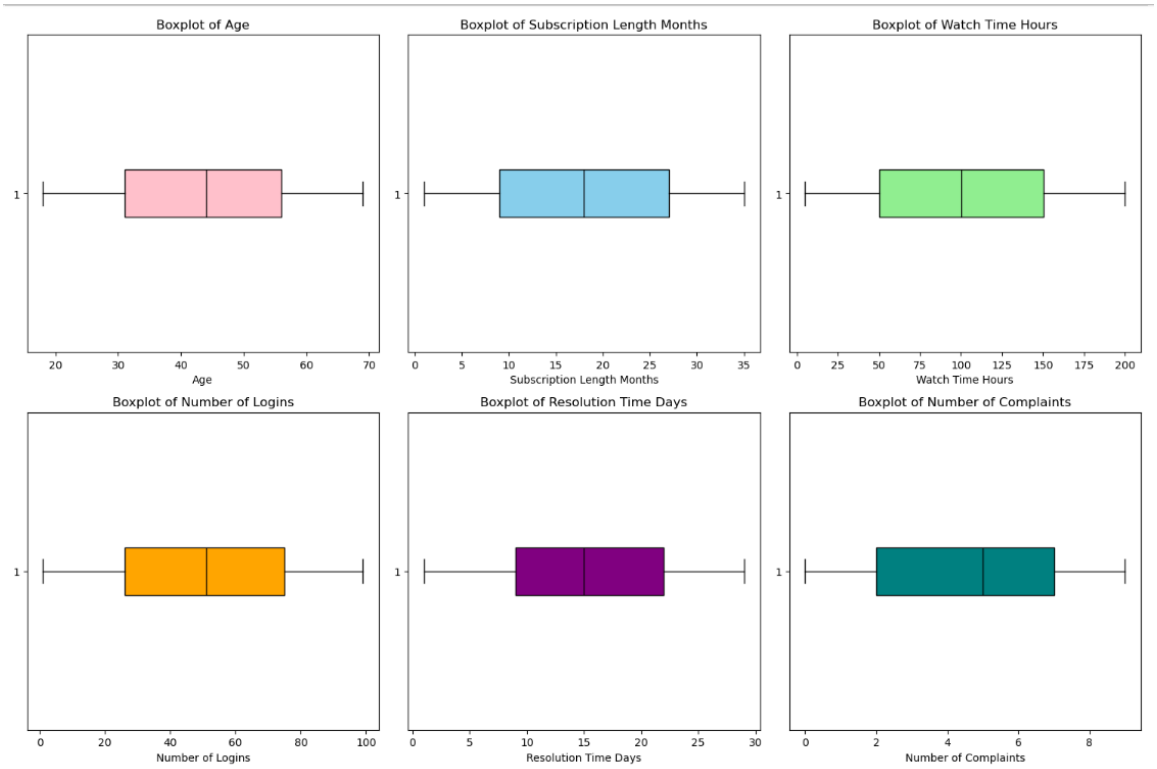
The dataset customer_churn.csv is loaded into Python using the pandas library.

```
   CustomerID  Age  Subscription_Length_Months  Watch_Time_Hours  \
0           1   56                          35         62.579266
1           2   69                          15        159.714415
2           3   46                          25         41.119547
3           4   32                          28        183.961735
4           5   60                          10         87.782848

   Number_of_Logins Preferred_Content_Type Membership_Type Payment_Method  \
0                73               TV Shows           Basic         PayPal
1                 1                 Sports           Basic    Credit Card
2                36                 Movies         Premium         PayPal
3                35                 Movies        Standard    Credit Card
4                66                 Movies        Standard  Bank Transfer

   Payment_Issues  Number_of_Complaints  Resolution_Time_Days  Churn
0               0                     7                     8      0
1               0                     7                    21      0
2               0                     5                    13      1
3               0                     0                    27      0
4               0                     7                    18      0
```

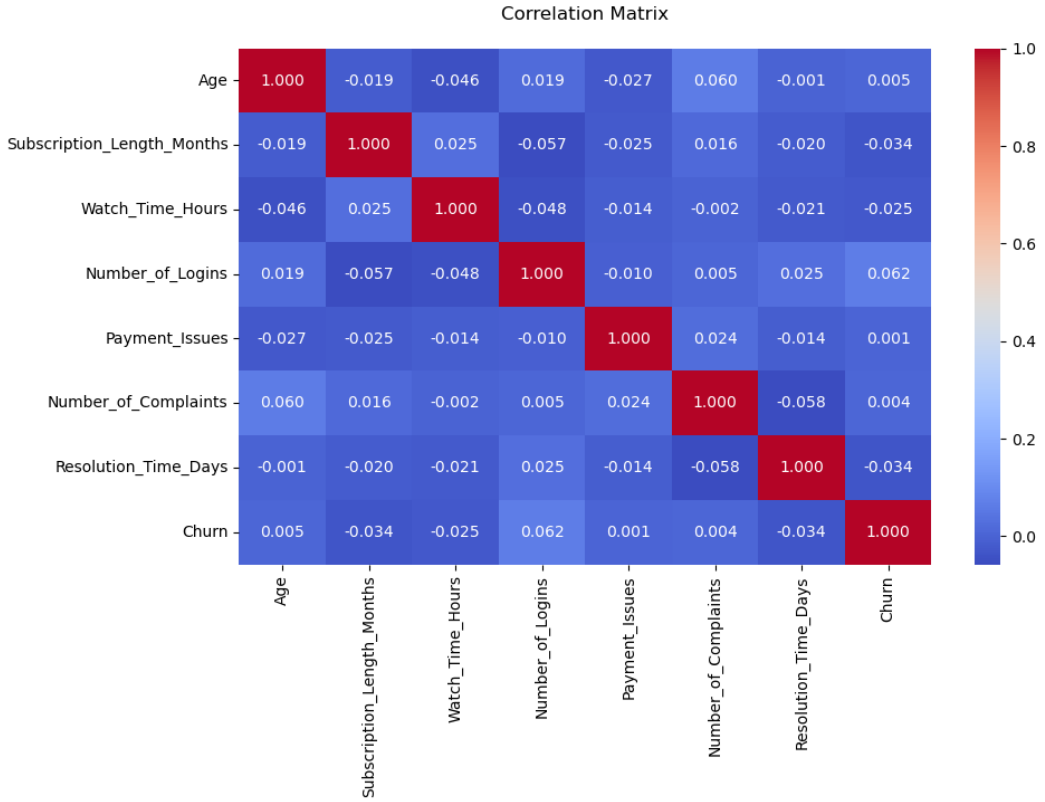
Exploratory Data Analysis (EDA)

- **Summary Statistics**: We used the describe() method to get summary statistics for numerical columns.
- **Missing Values**: We checked for missing values and found no missing values in the dataset.
- **Data Distributions**: We visualized the distributions of numerical features using histograms and box plots.

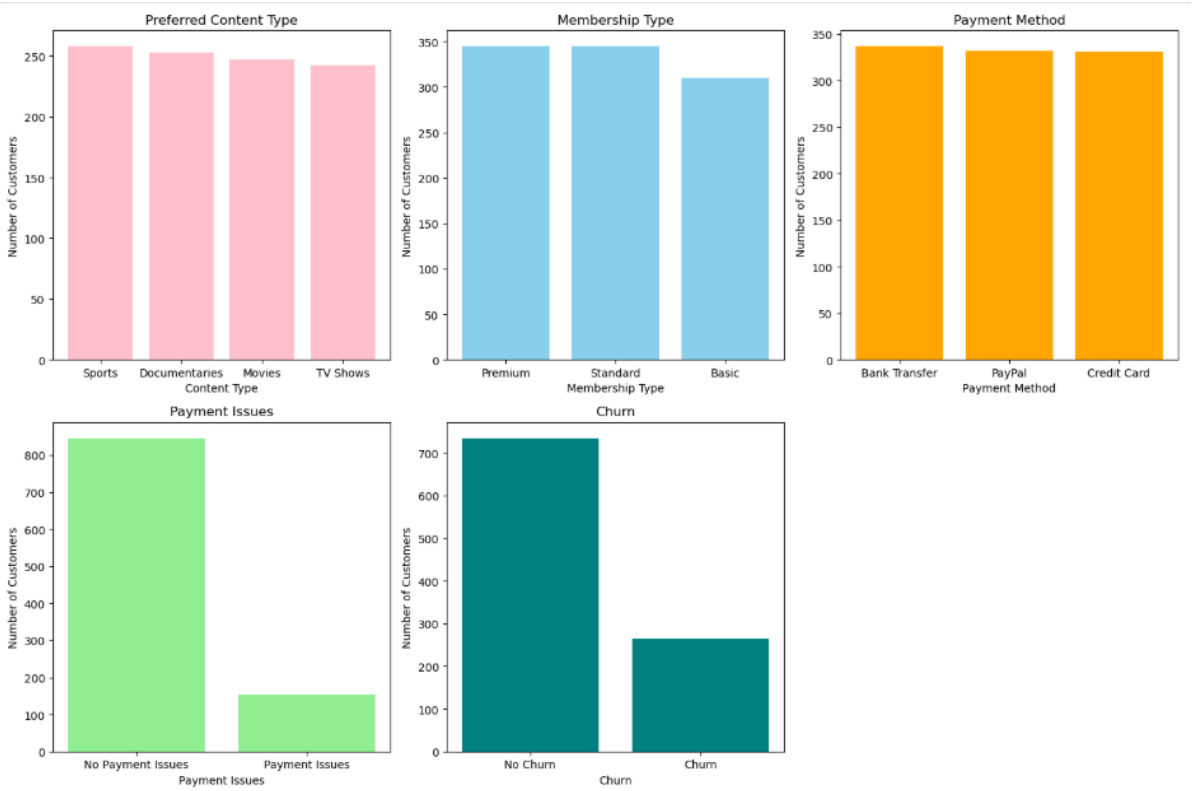




- **Correlations**: We check for correlations between variables using a correlation matrix and heatmap.



- **Bar Plots:** We visualized bar plots for Categorical features.



Task 2: Building a Decision Tree Classifier

Splitting the Dataset

We first split the dataset into training (70%) and testing (30%) sets while converting categorical variables into dummy variables. Since the dataset is small and imbalanced (majority class 0: 517, minority class 1: 183 in training), we needed to carefully handle class imbalance during model training.

Training the Initial Decision Tree

A Decision Tree Classifier was trained using default hyperparameters. The initial model achieved:

- Accuracy: 0.637
- Precision: 0.315
- Recall: 0.280
- F1 Score: 0.297

Hyperparameter Optimization with GridSearchCV

To improve performance, we tuned hyperparameters (max_depth, min_samples_split, min_samples_leaf) using GridSearchCV, optimizing for F1 Score (scoring='f1') instead of accuracy. Since the dataset is imbalanced, focusing on accuracy would favor the majority class. F1 Score balances Precision and Recall, ensuring better churn detection.

After tuning, the optimized Decision Tree achieved:

- Accuracy: 0.637 (similar to before)
- Precision: 0.358 (improved)
- Recall: 0.415 (improved)
- F1 Score: 0.384 (improved)

Although accuracy remained similar, Recall and F1 Score improved, meaning the model now correctly identifies more actual churn cases.

Task 3: Improving Performance with Random Forests

Initial Random Forest Model

We trained a Random Forest Classifier using default parameters. The model achieved:

- Accuracy: 0.717
- Precision: 0.200
- Recall: 0.012 (extremely low)
- F1 Score: 0.023

The model had high accuracy but very poor Recall, meaning it failed to detect actual churn cases.

Hyperparameter Tuning for Improvement

We optimized the model using GridSearchCV, adjusting n_estimators, max_depth, min_samples_split, and min_samples_leaf. After tuning, the model improved in detecting churn:

- Accuracy: 0.600 (slightly decreased)
- Precision: 0.279 (improved)
- Recall: 0.293 (improved significantly)
- F1 Score: 0.286 (improved)

Confusion Matrix (After Tuning)
[156 62]
[58 24]
More churn cases (24) are now correctly identified, improving Recall. However, accuracy dropped slightly due to better focus on the minority class.

Random Forest vs. Decision Tree Performance Comparison

Metric	Decision Tree (Tuned)	Random Forest (Tuned)	Best Model
Accuracy	0.637	0.600	Decision Tree
Precision	0.357	0.279	Decision Tree
Recall	0.415	0.293	Decision Tree
F1 Score	0.384	0.286	Decision Tree

Key Observations

- Decision Tree outperformed Random Forest in every metric, including Recall (0.415 vs. 0.293) and F1 Score (0.384 vs. 0.286).
- Random Forest had lower Accuracy (0.600) after tuning, as it focused more on improving Recall but still underperformed compared to the Decision Tree.
- Decision Tree handled the imbalanced data better, identifying more churn cases.

Analyze Feature Importance

To understand which features contribute the most to the model's predictions, we analyzed feature importance using the trained Random Forest classifier. we identified the key factors contributing to churn prediction.

Top Contributing Features

1. **Watch Time Hours (0.224)**

- The most significant factor in predicting churn.
- Customers who spend less time watching content are more likely to churn.
- **Actionable Insight:** Improve engagement by recommending personalized content to low-usage customers.

2. **Age (0.159) & Subscription Length (0.157)**

    - Older users and those with shorter subscription periods are more prone to churn.
    - **Actionable Insight:** Offer loyalty incentives to encourage longer subscriptions.

3. **Number of Complaints (0.112) & Resolution Time (0.111)**

    - Higher complaints & slow resolution times increase churn risk.
    - **Actionable Insight:** Improve customer support response times to retain users experiencing issues.

Mid-Level Influencing Factors

4. **Number of Logins (0.087)**

    - Less frequent logins indicate disengagement, potentially leading to churn.
    - **Actionable Insight:** Use reminders, notifications, and exclusive content to encourage logins.

5. **Preferred Content Type (0.044 - 0.009)**

    - Users preferring TV Shows (0.044) & Sports (0.020) are slightly more influenced in churn behavior.
    - **Actionable Insight:** Ensure strong content offerings in these categories to retain these customers.

Least Influential Features

6. **Payment Method (0.022 - 0.017)**

    - The choice of payment method (Credit Card, PayPal) has low impact on churn.
    - **Actionable Insight:** This indicates that financial convenience is not a major churn driver.

7. **Payment Issues (0.004)**

    - Surprisingly, payment issues are the least significant factor in predicting churn.
    - **Actionable Insight:** While not a strong predictor, ensuring smooth transactions can still improve user experience.

Task 4: Business Insights and Recommendations

Characteristics Contributing the Most to Customer Churn

Based on the feature importance analysis, the key factors influencing churn are:

- **Low Watch Time**: Customers who spend less time watching content are more likely to churn.
- **Subscription Length**: Short-term subscribers are at higher risk of churning.
- **Customer Complaints & Resolution Time**: High complaint rates and slow issue resolution contribute to dissatisfaction.
- **Age**: Older customers tend to have a higher risk of churn.

- **Low Login Activity**: Infrequent logins indicate disengagement, leading to churn.

Actionable Insights to Reduce Customer Churn

1. **Enhance Customer Engagement**: Target low-watch-time users with personalized content recommendations and exclusive offers.
2. **Improve Customer Support**: Reduce complaints and resolution time by streamlining issue resolution processes.
3. **Loyalty Programs**: Encourage long-term subscriptions through discounts, rewards, and personalized retention offers.
4. **Targeted Retention for Older Customers**: Develop engagement strategies tailored to older users, such as simplified navigation and curated content.

Three Concrete Business Strategies

1. **Proactive Customer Support**: Implement real-time issue resolution to address complaints before they lead to churn.
2. **Personalized Content Recommendations**: Use viewing history to suggest relevant content and increase engagement.
3. **Retention Campaigns for At-Risk Users**: Offer special discounts and promotions to customers showing disengagement patterns, especially older users.

Conclusion

We built and optimized Decision Tree and Random Forest models to predict customer churn. The insights revealed that watch time, subscription length, customer complaints, resolution time, age, and login activity are key churn drivers. By improving customer engagement, optimizing support, and offering retention incentives, StreamFlex can effectively reduce churn and increase customer loyalty.