Churn prediction in Theatre subscription

I`m analysing a business case in the entertainment industry, in particular I`m assessing what are the factors that lead a member *of an all you can watch cinema subscription* to churn or cancel their subscriptions.

My data set has almost 40,000 records of subscribers of a national theatre chain, the main Input variables on which build the algorithm are the variables below:

1. **TenureMonthly** = Indicates how many months the subscriber has his membership for in months
2. **Value** = Indicates how much the subscriber is paying per month
3. **MemberEquity** = this is possibly one of the most important variable as it determines if the subscriber is taking advantage of his subscription or not and it`s calculated as the difference between what the subscriber would pay at retail price for each show and what is paying now per show seen considering how much is paying monthly.
4. **Avg.TicketPrice** = indicates what on average the subscriber would pay for a single ticket if he wasn`t a member of the scheme
5. **MonthlyVisits** = how many times on average the member has visited the theatre since the start of the membership
6. **LocationCode** = Locations where subscribers use their membership
7. **VenueConditions** = State of the art of the venue, this variable indicates if the venue has been renovated or not and if it offers premium upgrades like recliners or premium seatings or Leisure activity and restaurants.
8. **TOTTransaction** = This variable indicates how many transactions a member made at any concession stand or any restaurant.
9. **Gifted** = this variable indicates if the membership has been purchased by the specific member or has been gifted by someone else.
10. **Distance** = Indicates the distance between the theatre and the subscriber`s home address.
11. **Date of Birth** = Age.
12. **Gender** = Male, Female.
13. **Status** = Active, Churned.

The Output variable is the Status which indicates if the member has churned the scheme or is still active.

Design of a Decision Tree explaining who is most likely to churn

These are the steps to build a Decision Tree to predict the subscribers most likely to churn

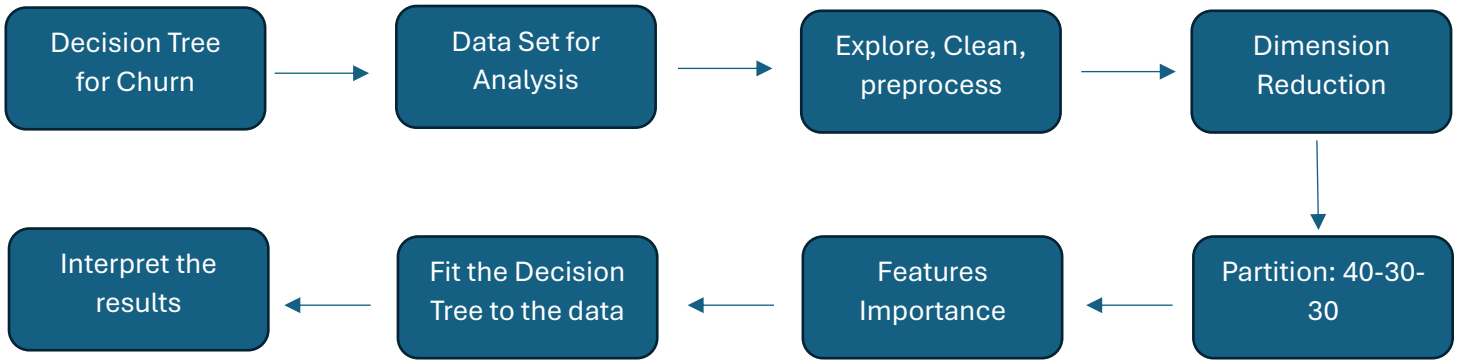
1. I want to drop the variable FiscalYear_New.
2. I want to explore the dataset and highlight the outliers.
3. I want to transform the VenueConditions in numerical, by transforming the categories Poor, Average and Good in 1,2,3 respectively.
4. I want to transform as well the categorical variables Gifted in a binary variable where 0 = Self and 1 = Gifted.

Outliers

In terms of Outliers I noticed almost 400 records with TenureMonthly = 1 with a number of visits too high for that tenure, as a result I decided to drop the rows for which TenureMonthly = 1 as I considered those observations not correct.

Dimensions Reduction and Decision Tree implementation

In the Dimension reduction I selected the features of the members who are more likely to churn, dropping variables like: TOTCost, TicketRevenue and Value as correlated with variables I decided to keep in the model (MembershipValue, Avg.TicketPrice and MembershipEquity), I dropped Gender as not populated correctly, then I partitioned the data set into Training, Validation and Test.



Features importance: Analysis of the importance of the main input variables and Validation and Test set classification report

After having transformed some of the categorical variables in numerical I proceeded to analyse the importance of my input variables and I assessed the Validation Set and Test Set Classification Reports

Both the Validation Set and Test Set Classification Reports show similar performance metrics:

Accuracy: The accuracy is around 72% on the validation set and 73% on the test set. This indicates that the model correctly predicts the status (Churned or Active) for approximately 72-73% of the subscribers in both datasets.

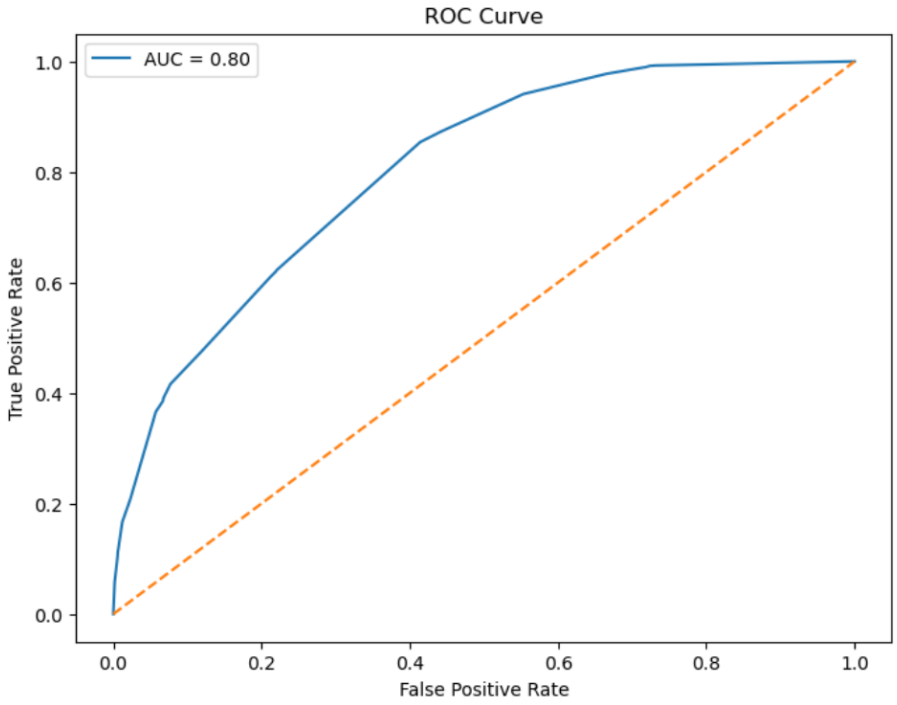
Precision: For class 0 (Churned), the precision is 76%. This means that when the model predicts a subscriber will churn, it is correct 76% of the times. For class 1 (Active), the precision is 69%. This means that when the model predicts a subscriber will remain active, it is correct 69% of the time.

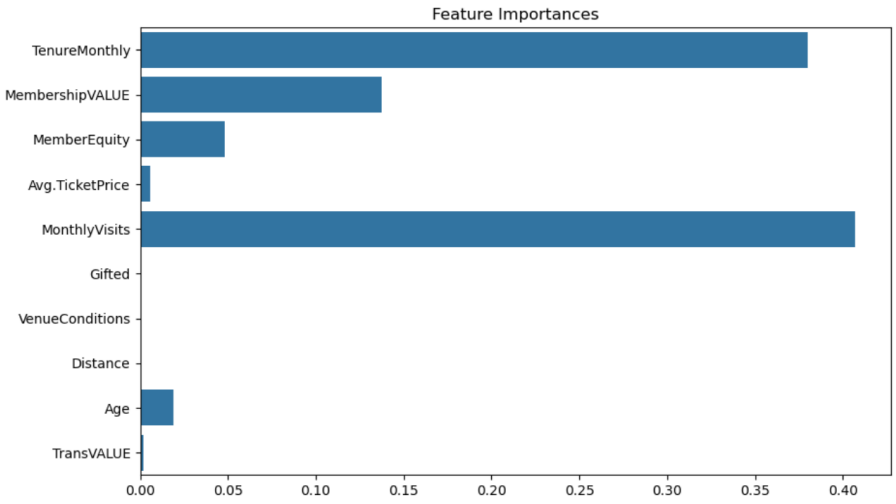
Recall: For class 0 (Churned), the recall is 78%. This means that the model identifies 78% of all actual churned subscribers. For class 1 (Active), the recall is 67%. This means that the model identifies 67% of all actual active subscribers.

F1-score: The F1-score balances precision and recall. The F1-scores are 0.77 for the churned class and 0.68 for the active class.

ROC Curve Analysis

- **AUC (Area Under the Curve):** The AUC is 0.8. This indicates that the model has a good ability to distinguish between churned and active subscribers. An AUC of 0.8 suggests a reasonable level of discrimination, better than random chance (0.5) but with room for improvement to reach a near-perfect score of 1.





The most important predictors from an input variable perspective seem to be: MonthlyVisits, TenureMonthly, Value, TransVALUE, MemberEquity, Age, Avg.TicketPrice, whereas the distance between the subscriber`s house address and the theatre alongside the venue conditions don`t seem very important.

Decision Tree analysis

The decision tree aims to predict theatre`s subscribers churn, classifying customers as either "Churned" or "Active." It uses several features, explained in the chart above.

Root Node

1. **Feature:** MonthlyVisits
2. **Condition:** MonthlyVisits ≤ 1.5
3. **gini:** 0.487
4. **samples:** 15292
5. **value:** [8872, 6420] (8872 Churned, 6420 Active)
6. **class:** Churned

Interpretation: The root node splits the data based on whether the number of monthly visits is less than or equal to 1.5. Initially, the dataset contains 15450 samples, with a higher proportion of churned customers (8993) compared to active customers (6457).

Key Splits and Branches

Left Branch (MonthlyVisits >=1.5):

Node: TenureMonthly ≤ 19.5

gini = 0.496

samples = 9743

value = [4438, 5289]

class = Active*:* For customers with 1.5 or more monthly visits, the tree then considers "TenureMonthly." If tenure is less than or equal to 19.5 months, the majority of customers are classified as "Active."

Further Split: MembershipValue <= £14.47

gini = 0.499

samples = 6956

value = [3641, 3315]

class = Churned *Interpretation:* If Membership value is less than £14, the majority is classified as "Churned, although the Gini Index is high and the impurity of the node as well.

Right Branch (MonthlyVisits <= 1.5):

Node: TenureMonthly ≤ 70.5

gini = 0.326

samples = 5692

value = [4524, 1168]

class = Churned *Interpretation:* For customers with less than 1.5 monthly visits, the tree considers "TenureMonthly." If tenure is less than or equal to 70.5 months, the majority of customers are classified as "Churned." Gini impurity at 0.326 gives us an indication of the impurity of the node, only 20% of the subscribers are classified incorrectly.

Further Split: MembershipValue <= £12.59

gini = 0.252

samples = 5032

value = [4288, 744]

class = Churned *Interpretation:* If MembershipVALUE is less than or equal to £12.59, the majority are classified as "Churned." Impurity of the Gini index drops to 0.25 providing further evidence of purity in the node relative to the class Churned.

Further Split: TenureMonthly <= 62.5 (Less than 5 years and 2 months)

gini = 0.115

samples = 2992

value = [2809, 183]

class = Churned *Interpretation:* there is another split on TenureMonthly, in particular if it is less than or equal to 63 months, the majority are classified as "Churned." Impurity of the Gini index drops to 0.115 providing further evidence of purity in the node relative to the class Churned.

Insights

- **Monthly Visits Impact:** The number of monthly visits is a strong initial indicator of churn. Customers with fewer visits are more likely to churn. This is the most significant sign of disengagement.
- **Tenure Influence:** Tenure plays a significant role, especially when combined with monthly visit data. Shorter tenures often correlate with higher churn rates.
- **MembershipVALUE:** This is an important feature in predicting the behaviour of the subscriber, in fact it represents the monthly fee is paying minus the costs of each show (screen player's rights) and it represents the value for the theatre operator
- **MemberEquity:** Member equity is considered in one of the branches, suggesting it has some predictive power, particularly for identifying active customers, this is an important measure when it comes to assess if the subscriber is squeezing value out of his membership.
- **Gini Impurity:** The Gini impurity values indicate the homogeneity of the nodes. Lower Gini values in the leaf nodes suggest more reliable classifications.

Possible applications of this Machine Learning model

The model is suitable to be applied for churn detection in subscription businesses, with adaptability for entertainment: theatres, cinemas and leisure parks.