Leadership and Consulting

FITNESS CENTER EXPLORATORY DATA ANALYSIS

By: Suyog Mahale

INDEX

THE BUSINESS PROBLEM

- Fitness center XYZ has recently been losing quite a few customers (Churn). While customer acquisition has been going well, customer retention has been a problem. We need to analyze exactly why customers are churning and get to the bottom of the issue.

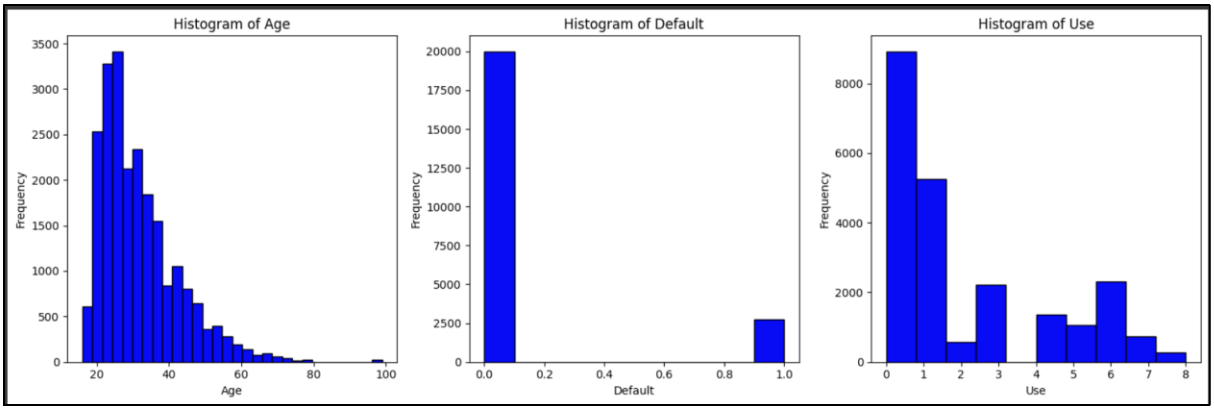
- The goals of our analyses are:
    - To understand which model we should select to identify which members are most likely to churn.
    - Identify the demographics of customers most at risk to churn.
    - Draft a systematic approach to tackle the issue of churning by coming up with a plan of action.

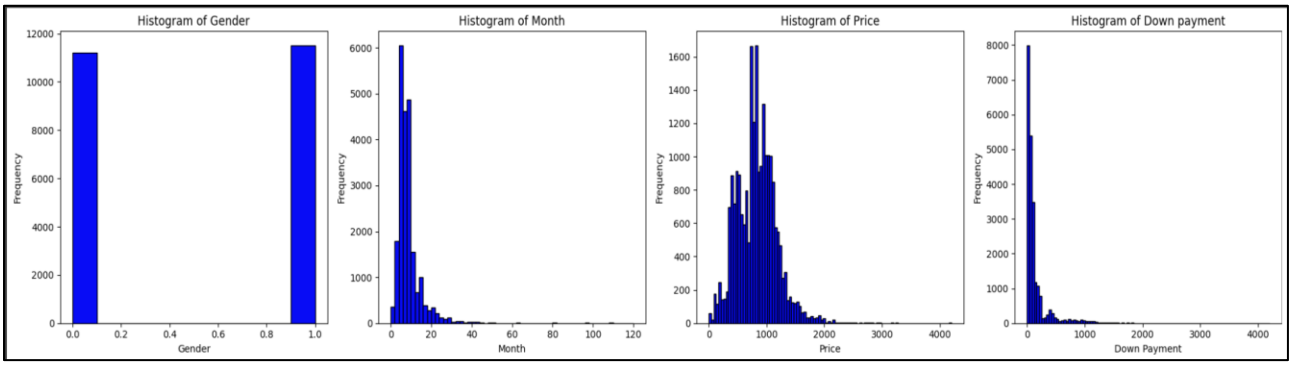
MODEL SELECTION AND DATA

From the data we have been given, there are a few important variables.

1. Price – The price of the membership purchased by members (Numerical)
2. Downpayment – The initial payment made by members. (Numerical)
3. Monthdue – The duration of the membership (Numerical)
4. Use – the usage frequency of the member. (Categorical)
5. Age – The age of the member. (Numerical)
6. Gender (Categorical)
7. Default – Whether the member has churned or not. (Categorical)

On analyzing the frequency counts of the variables, we observe the following:



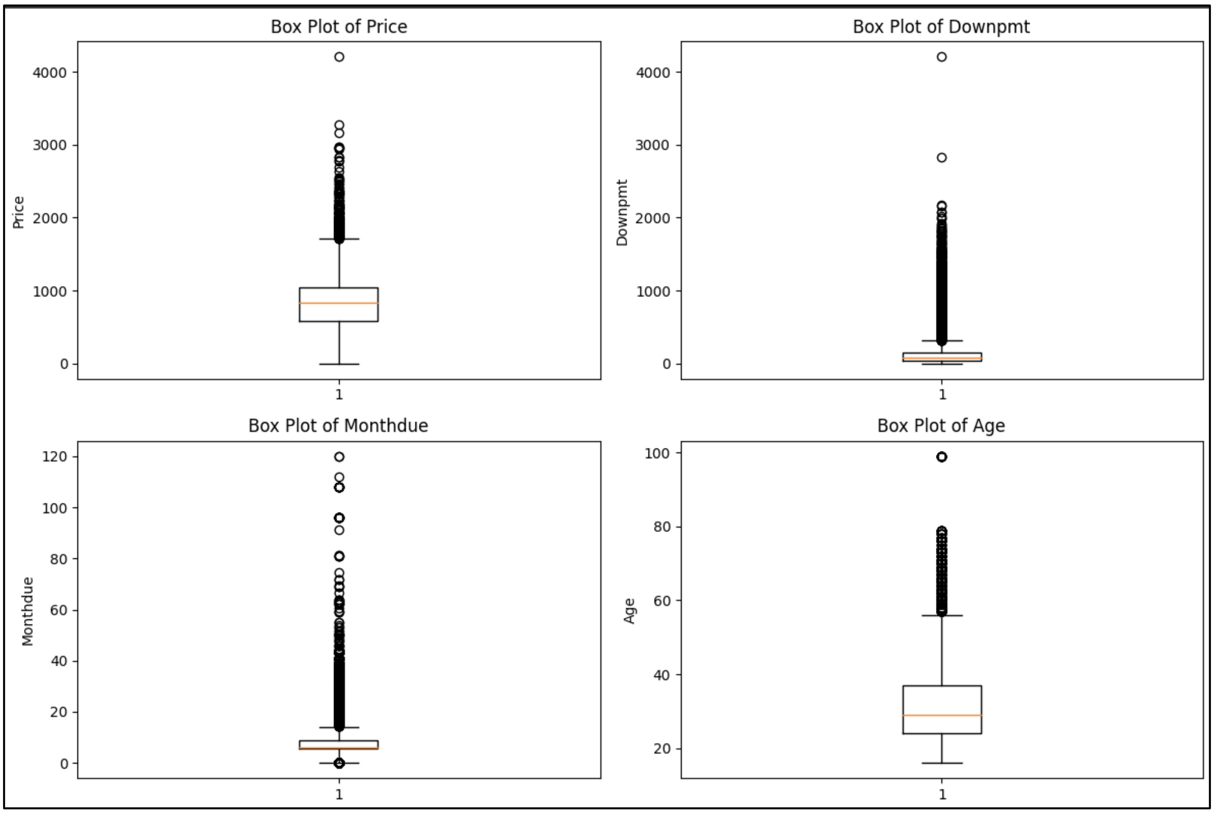


- Majority of Fitness Center XYZ's members are Young to Middle-Aged adults between 20-40 years old.
- There are equal number of men and women joining Fitness XYZ
- Most members have a membership duration of about 6-12 months.
- The churn rate of members from the given data is approximately 12.5%.

Outliers

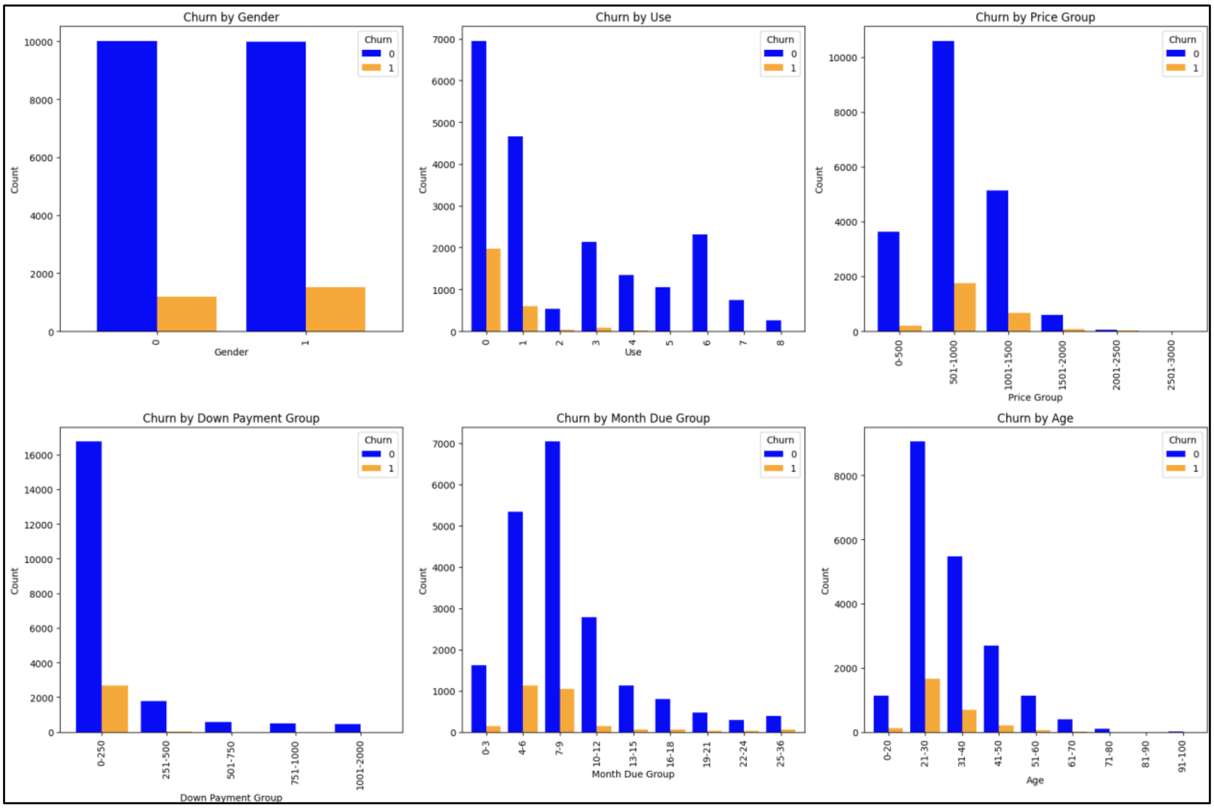
There appear to be quite a few outliers for each of the variables in the data that we've been given.

- The ages of some records are between 0-16 which appear to be too young to have a membership at Fitness XYZ. For our analyses, we have cleaned up all records with age < 16.
- Most variables, as we can see from the above histograms, have extremities which are not practical. Such as memberships > 60 Months, Price above 300, Down Payment above 1000, and other such abnormalities in the data. For our analysis, we will only consider data between the 5th percentile and 95th percentile of the data range and drop the data with extremities to handle the outliers.



Preliminary Analysis

We have conducted a preliminary analysis of the given data by analyzing each of the given variables and their relationship with member churn. From the preliminary analysis we observe the following:



In the above graphs, blue bars refer to members who have not churned and yellow bars refer to churned members.

- We can observe from the above that members who have less use frequency are more likely to churn which is implicit because if members have no use at the fitness center, they are unlikely to continue their membership.
- In terms of finances, we see that churn rate is higher in members who have a middle range priced membership. Also, the lower the down payment, the more likely a member is to churn, which also indicates commitment.
- Similarly, we can observe that the shorter a member's membership term, the more likely they are to discontinue their membership, thus increasing churn.
- Finally, on observing the age groups, we see that ages 20-40 have the highest churn however this group also has the highest number of loyal members. The churn rate between men and women is approximately the same.

Model Selection

To conduct a more detailed analysis to identify the factors impacting churn, we need to create a machine learning model with the data that we have been provided. The data for our machine learning model is primarily categorical and numerical data. Based on this, the following models are most suitable for analysis:

1. Logistic Regression

- Logistic regression is a simple, interpretable model which is good for binary classification problems such as predicting churn. It models the relationship between one or more independent variables (features) and the binary dependent variable (target) using the logistic function (also known as the sigmoid function). Logistic regression outputs probabilities that a given input belongs to a specific class, typically 0 or 1.
- Logistic regression is simple to train, and it is easily interpretable. However, this model has many issues in relation to our data. Logistic regression is sensitive to outliers and may not handle them well. It also does not handle complex relations and struggles with imbalanced data. It also assumes a linear relationship between the variables. All these factors make it not the most suitable model for our analysis.

2. Random Forest
    - Random Forest is a powerful ensemble method that can handle both numerical and categorical data. It can capture non-linear relationships between the features and the target (churn). It also provides feature importance, which can be useful for identifying the key drivers of churn.
    - Random forest can handle outliers, class imbalances, and it is flexible and scalable. Based on the data that we have been given, random forest seems like the most suitable model for analysis.
        - There are a lot of noisy variables such as payment method and use which. Random forest can handle the noise.
        - Random forest will give us feature importance scores which will help us identify which features are the strongest predictor of churn.
        - Since there a lot of nonlinear relationships in our data (price, downpayment, use, months due) random forest will be suitable for capturing these relationships.

Conclusion

We will start off our analysis with a random forest model which will help us give a clear idea of the factors most impacting churn. We will calculate key metrics such a precision, F1 Score, and ROC-AUC to get detailed information about our data.

Next, we will create a logistic regression model which will be easily interpretable for our stakeholders and help us explain basic relationships between the data.

We will compare the results from the two models and from the insights we have gained, draft a plan of action for Fitness XYZ for the next 6-12 months.

QUESTIONS

Considering the data that we have been provided and the information we have of Fitness XYZ, there are a few questions that we would like to get clarified before we proceed with our analysis.

1. First, we would like to get clarifications about the data. There are quite a few records with ages below 16 and ages all the way down to 0. We would like to ask whether these are corrupted records or is the data just missing. Should we consider this data for our analysis?

2. We would like to gain a deeper insight on member data collection practices at Fitness XYZ. What exact data is collected at the time of member sign-up? How does Fitness XYZ identify it members (Unique Member identification, Membership code)? Which data can legally be provided to us for our analysis?

3. Is it possible to arrange visits to Fitness XYZ locations to gain an understanding of operations and experience at Fitness XYZ? We would like to talk to the front desk and store managers about customer frequency, which time of day is the busiest, which day of the week is the busiest, etc.

4. Are marketing practices at Fitness XYZ focused on new customer acquisition or customer retention? What are the current ongoing marketing endeavors/offers in place to incentivize members to continue membership? What is the Fitness XYZ's marketing budget?

SUMMARY

From our preliminary analysis, we can see that there are multiple factors which could affect churn at XYZ. To conduct a more detailed analysis, we will be using an optimized random forest regression approach to model the data. This will give us a deeper understanding of what factors are driving churn and what we can do to circumvent these issues. Based on the inferences we can implement the following:

1. Create targeted campaigns for member groups which are most likely to churn.
2. Analyze the loss to benefit ratio of marketing practices. If Fitness XYZ is getting more loyal members rather than short term churning members, we can keep focusing on the loyal customers.
3. Identify the cause of member churn and target our focus on most frequent causes of churn.
4. Track retention metrics and refine strategies accordingly.

By identifying key factors and targeting churn, we aim to reduce churn at Fitness XYZ by 10%-15% in the next 6 months.