Assignment 6

Predictive Model Exercise

Submitted by
Swathi Ganesan
12372237
swathiganesan@uchicago.edu

The problem statement – Fitness Center case study

The consulting team's approach addressing the Fitness Club's increasing churn rate is to design and build a customer segmentation model via cluster analysis that will help us identify customer archetypes that are highly likely to churn in order to help the fitness club achieve a more effective customer marketing strategy by personalization so that they can curb churn and increase customer retention period.

Step by step Predictive Modelling Process

a. Understanding the Data and Data Cleaning

We start the process with importing the relevant libraries and loading the Fitness Club dataset. After looking at a snapshot of the data using the head() method and understanding the datatypes of the columns using the info() method I was able to discern the following

Numerical features:

price
downpmt
monthdue
age

Categorical features:

enrolldt
pmttype
use
gender
default

A quick check for null or missing values showed that there were no null values. However, the data seemed to have 28 duplicate entries and we proceed by dropping the duplicate entries in the table.

We obtain the descriptive statistics on the numerical features of the data to understand the spread of data using count, mean, standard deviation, minimum, maximum and the percentiles.

We add descriptive columns for gender and churn columns (with the assumption that Gender 1 – Female, Gender 0 – Male, Churn – 1 and No Churn - 0) as they are categorical features with numeric values in order to improve readability of the data. We understand the distribution of the categorical features using value_counts().

We plot some basic visualization on the data to understand the data and the features better before proceeding to modelling. Plotting histograms as they are good indicators of the data distribution and spread along with helping us identify outliers and high-leverage points in the dataset.
We plot histograms on the entire dataset and the customer data for churned customers only. On comparing the two we can identify the following:

- Female customers are more likely to churn
- Customers putting in lower down payment are more highly likely to churn
- We must focus on retaining customers between the ages of 20 to 40
- Customers who visit the gym less frequently should be encouraged to work out more often to retain them

We plot correlation matrix or the correlation heat map to find potential relationships between performance metrics and success statistics and to understand the strength of these relationships. We can see a slightly good correlation between down payment and price but as that does not really help us in the churn analysis, we can ignore it.

b. Data Pre-processing and Splitting

From the descriptive statistics we can identify that age has min value 0 and max value 99. Logically as we are dealing with Fitness club members, we can safely drop customer data having age below 16 and greater than 80.

In case we want to understand the seasonality in the customer enroll dates, we have converted the enrolldt column to datetime type.
We can notice that we do not have complete information for the year 2021 as data is available for only 3 months and the payment method Cash has been removed for the year 2021. In order to avoid bias in out model prediction and for consistency we can drop customer data for the year 2021.

To input into the model we need numeric datatypes, to achieve this we can do one-hot encoding or dummy/ indicator variables for the payment type column and the usage frequency columns. In our case we have created corresponding indicator variables for the same. The data is now ready for the model on selecting the features that we would like to input to the model, in our case : *'price', 'downpmt', 'monthdue', 'age', 'gender', 'default', 'pmttype_Cash', 'pmttype_Cheque', 'pmttype_CreditCard', 'pmttype_DirectDebit', 'use_0', 'use_1', 'use_2', 'use_3', 'use_4', 'use_5', 'use_6', 'use_7', 'use_8'*

We need a train-test split on the data to estimate the performance of our Logistic Regression model in predicting churn. We make a train : test split in the ratio 80:20. We specify the stratify parameter as we want to preserve the same proportions of examples in each class as observed in the original dataset.

c. Model Building

We are building a Logistic Regression model to predict churn/retention using the LogisticRegression class available in the sklearn.linear_model package. We fit the mode using our train data (X_train and y_train) and we predict the y_pred values using the X_test input.

We would now need to evaluate our model by comparing the y_pred and y_test values.

d. Model Outcomes and Validation

We can evaluate the model using the following methods. Each method has its benefit, and we can explain it as follows:

- **Confusion matrix**

  Confusion matrix is a method used to summarize classification algorithm on set of test data for which the true values are previously known. Sometimes it also refers as error matrix. It gives us the following information.

  True positive [0,0]: TP means model predicted yes and correct answer for that is also yes

  True negative [1,1]: TN means model predicted no and correct answer for that is also no

  False positive [0,1]: FP means model predicted yes but actual answer is no

False negative [1,0]: FN means model predicted no but actual answer is yes

```
The confusion matrix is :
                      [[3182  148]
                      [ 272  242]]
```

From the confusion matrix we can understand that we have 3182 True Positives and 242 True Negatives. To interpret the model performance from the **Confusion Matrix** we can compute Accuracy as follows:

Accuracy = (TP + TN /Total) = 89.07%

- **Classification report**

  A Classification report is used to measure the quality of predictions from a classification algorithm. It gives us the precision, recall, f1-score, and support metrics which are explained below

  Precision - Accuracy of positive predictions: TP/ (TP + FP)

  Recall - Percent of the positive cases identified correctly: TP/ (TP+FN)

  F-1 score - 2*(Recall * Precision) / (Recall + Precision)

  The F1 score is a weighted harmonic mean of precision and recall. Generally, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, **the weighted average of F1 should be used to compare classifier models, not global accuracy**.

  Hence, we can identify from the **Classification report** that model accuracy is **88%**

- **ROC curve**

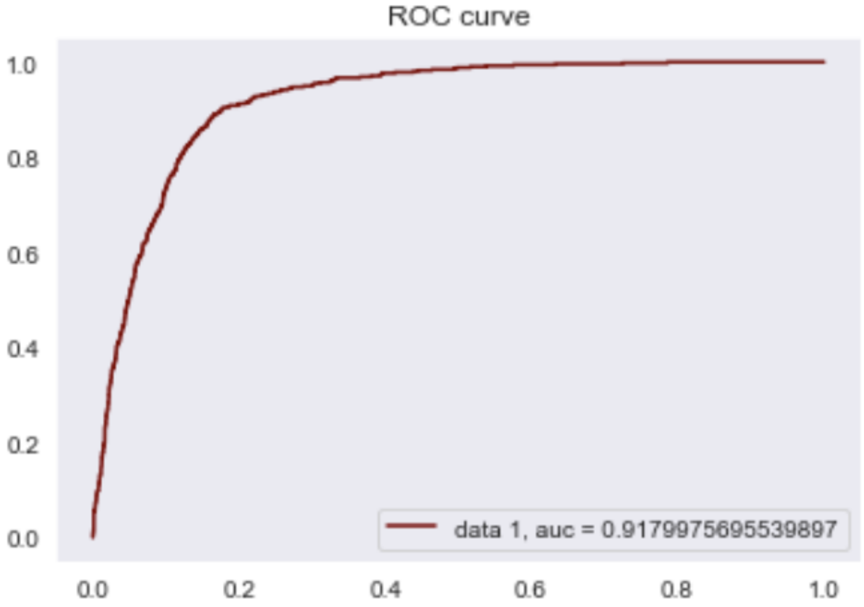
  ROC- Receiver operating characteristic curve will help to summarize model's performance by calculating trade-offs between TP rate and FN rate and it will plot these 2 parameters. To classify this term AUC (Area under the curve) is introduced which gives summary of ROC curve.

  The higher the AUC, the better the performance of classifier.
  We can interpret the AUC values as follows:

1) If AUC =0 then classifier is predicting all the positive as negative and negative as positive.

2) If 0.5< AUC < 1 means classifier will distinguish the positive class value from negative class value because it is finding a greater number of TP and TN compared to FP and FN.

3) If AUC = 0.5 it means classifier is not able to distinguish between positive and negative values.



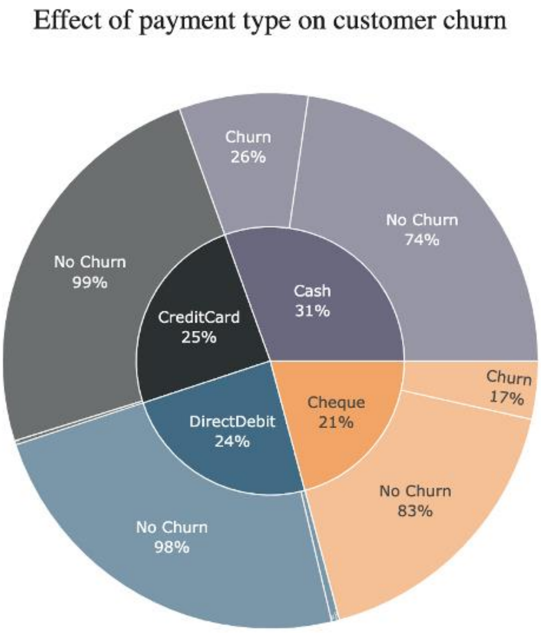
The ROC curve has **Area Under the Curve = 91.79%** since the auc value falls between 0.5 and 1 we can determine that we have built a good model.

We perform feature selection to narrow the set of features to those most relevant to the machine learning model. The feature selection objective in machine learning identifies the most helpful group of features that can be used to build useful models by eliminating redundant and irrelevant features.

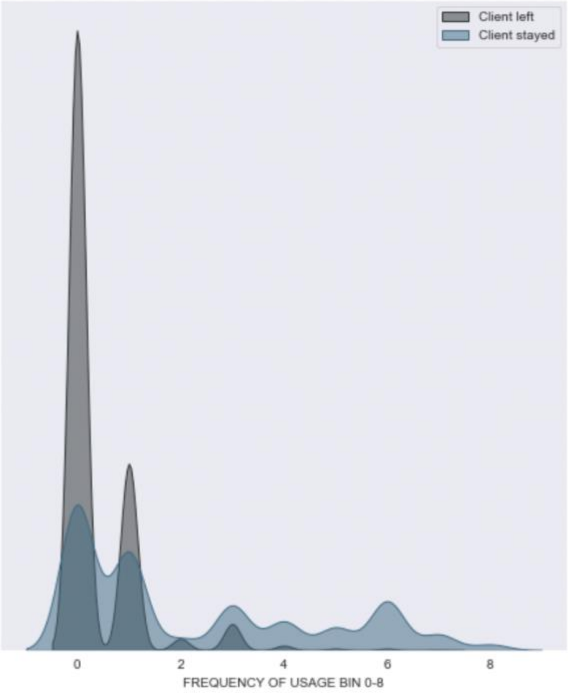
On applying the feature selection, we can identify that **payment type and frequency of use** are the most relevant features and can be identified as the key drivers of the business.

We can conclude that our model performance is satisfactory as we have a very good accuracy and auc as seen above and our findings are in alignment with our initial EDA where we identified the *pmttype and use* features as the more insightful data points. Recalling out findings from the EDA as follows.

Visuals from the EDA that illustrate insights about the data and business



Effect of payment type on customer churn



From the images we can understand that customers using Direct Debit and Cash as the payment methods are more likely to churn. We also learn that regular customers have high retention rates when compared to irregular customers.

Hence, we can help the Fitness Club incentivize customers using the Credit Card payment method in an attempt to encourage more Credit Card payments and help retain customers. Additionally, the Club can motivate the irregular customers to visit more frequently by introducing challenges with perks and by keeping their customers engaged with new weekly workout routines like Pilates, Yoga etc. Change in customer visit trends can easily be translated to higher retention.