# Assignment 2: Juan Escalada

## Introduction

In this report, we will be using customer churn data to predict whether a given customer will churn (stop using the service) given their current attributes.

Some of the attributes that we will be exploring include:

* Account: Account Age (months), Customer ID, Gender
* Devices: Device Registered , Multi-Device Access
* Payments: Monthly Charges, Paperless Billing, Payment Method, Subscription Type, Total Charges
* Preferences: Genre Preference, Parental Control, Subtitles Enabled
* Usage Statistics: Average Viewing Duration, Content Downloads Per Month, Content Type, Support Tickets Per Month, User Rating, Viewing Hours Per Week, Watchlist Size

We will begin by exploring the data in the dataset, to get an idea of what the typical customer looks like, as well as summarizing any patterns and correlations that may arise.

After that, we will create 3 models based on our analysis, and select the best model.

Our best model, along with our other models have a few variables which were highly significant:

Account Age, Monthly Charges, Viewing Hours per Week, Average Viewing Duration, Content Downloads pre Month, Support Tickets per Month, Payment Method (Mailed Check), Paperless Billing (No), Content Type (TV Shows) and Multi-device Access (No).

## Exploring the Data

We can start our data exploration by visualizing the typical ranges for various numerical parameters based on whether the user has churned or not. Observe the following boxplots (Fig. 1, 2, 3, 4):

A diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generated

Fig. 2: Monthly Charges, Grouped by Churn

Fig. 1: Account Ages, grouped by Churn

Here, we can see that accounts that churn are typically around 24 months newer than accounts that don’t. This indicates that user loyalty increases over time. On the other hand, monthly charges are slightly higher for accounts that churn, perhaps pointing to an issue with pricing strategies.

A diagram of a box plot

Description automatically generatedA graph with blue and green lines

Description automatically generated

Fig. 4: Support Tickets per month, grouped by Churn

Fig. 3: Viewing Hours per week, grouped by Churn

We can see that users that churn tend not to spend as much time using the platform, as the average viewing times are about 5-6 hours less, even for avid users. Meanwhile, users who churn seem to make on average 1 extra support ticket per month, perhaps indicating dissatisfaction with the service as their main reason to cancel their service.

It’s also interesting to look at the churn rates for different groups. This will allow us to identify the categorical variables that influence the churn outcome. We can see some rather unexpected results after calculating the churn rates for certain features, such as the user’s Genre Preference and the type of content they like to view (Fig. 5, 6):

A graph of a bar chart

Description automatically generated

Fig. 5: Churn Rates by Genre Preference

A graph of blue rectangular bars

Description automatically generated with medium confidence

Fig. 6: Churn Rates grouped by Subscription and Content Types

Depending on the user’s favourite Genre, the churn rates can vary from 16.56% to 19.28%. In other words, Comedy enjoyers are 16% more likely to churn versus Action enjoyers. This is a rather surprising finding, especially considering that the average churn rate is 18.13%.

In Fig. 6, we plot the churn rate of all combinations of Subscription and Content Types. Possible Subscription Types are: Basic, Standard and Premium. Possible Content Types are: Movies, Shows and Both. We found that the churn rates for Premium users that view exclusively TV Shows are 2.5% lower than the mean, which is very significant. On the other hand, Basic users that view both TV Shows and Movies have a churn rate almost 2.5% higher than the norm.

Based on our previous boxplots, we decided to separate the account ages into bins, by dividing the account age by 12 and rounding it down. This gives us the number of years that the account has been active. This is by far the most significant indicator of whether a user will churn (Fig. 7):

A graph with blue and gray bars

Description automatically generated

Fig. 7: Churn Rates grouped by Account Age (years)

New users (<12 months active), have a very high churn rate of 31.18%, compared to an abysmal 7.68% churn rate for users who have been active for over 9 years. This means that churn rates fall as the users become more loyal. Notably, the churn rate falls below the average after the 5-year mark, showing that there is a high risk of losing users during the first 4 years.

Furthermore, we analyzed the churn rates depending on payment method. Inconvenient payment methods such as checks have churn rates 3% higher compared to credit cards (Fig. 8). This means they are almost 20% more likely to churn:

A graph of a number of blue bars

Description automatically generated with medium confidence

Fig. 8: Churn Rates grouped by Payment Method

Finally, we analyzed the churn rates based on each account’s User Rating. To do so, we separated the account into 4 bins: Average rating 1-2, 2-3, 3-4 and 4-5. This represents the average rating that the users gave to the content they watched. Surprisingly, as the churn rates go down as the user rating goes down. In other words, users that were more ruthless in their ratings, were less likely to cancel their subscription (Fig. 9). High raters were almost 15% more likely to cancel their subscription that low raters.

A graph of blue rectangular bars

Description automatically generated with medium confidence

Fig. 9: Churn Rates grouped by User Rating

## Developing Models

We developed various logistic regression models to predict whether a given user will churn, namely, cancel their subscription. Model 1 includes only the highly correlated factors we identified before, categorized using some basic bins. Model 2 uses recursive feature extraction (RFE) to identify the best predictors among the binned variables and the pre-existing ones. Finally, model 3 makes use of every single variable, to identify any potential predictor we may have missed.

Models 1 and 3 make use of SCUT (SMOTE and Clustered Undersampling Technique) which allows interpolating values for the churned users which are quite few compared to the non-churned ones. On the other hand, model 2 uses JUST (Juan’s UnderSampling Technique) an experimental technique for handling imbalanced data by selecting a random sample among the class with the most total samples.

The bins in all the models were selected conveniently according to the data we analyzed previously. Ages have been grouped according to years rather than months. User Ratings have been rounded to the nearest 0.1 to make them easier to work with, and then binned into 4 groups (1 to 2, 2 to 3, 3 to 4 and 4 to 5) which are representative of all the available samples.

### Model 1

This is a simple model which uses some of the basic influential features we identified previously: Account Age, Viewing Hours per Week, Genre Preference and User Rating. As performance by itself was quite low, binning was required to improve the model’s viability, as well as the strength of its main predictor, Account Age.

We implemented scaling to boost model performance. On top of that, invalid values have been imputed using the median and standard deviation which would be expected depending on whether the user has churned or not.

### Model 2

This is another relatively simple model, which starts by generating dummy variables for every single categorical feature. The 10 most relevant features are selected using Recursive Feature Extraction. Imputing via K-nearest Neighbours was also employed to fill in the missing values and increase performance, however, the production code does not apply this imputation algorithm, choosing instead to use a randomized normal distribution.

This model makes use of JUST for data balancing rather than oversampling. This allows it to process the entirety of the samples in a very short time, which makes this model scalable. It shouldn’t have trouble processing millions of rows, which would take weeks to process using SMOTE/SCUT.

### Model 3

This is a model that uses all of the existing variables indiscriminately. The processing speed is relatively low, therefore sampling the original dataset (taking only about 30% of the original data) was required to keep processing times at a reasonable enough pace.

## Evaluating the Models

To evaluate the models, we use various metrics which are commonly used in machine learning, particularly for assessing Logistic Regression models: Accuracy, Precision, Recall, AUC and F1. Our primary guiding metric is F1, which is like a weighted average of the Accuracy, Precision and Recall.

Each model is cross-validated 5 times to achieve more accurate results for each of these metrics. Each metric has a mean and a standard deviation between all the generated cross-validated models.

The parameters for each model are shown in the diagram below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mean (SD) | Parameters | Accuracy | Precision | Recall | F1 |
| Model 1 | 22 | 62.6% (0.62%) | 27.4% (0.44%) | 64.0% (0.81%) | 38.3% (0.38%) |
| Model 2 | 10 | 67.9% (0.7%) | 67.6% (1.2%) | 69.2% (0.52%) | 68.4% (0.78%) |
| Model 3 | 52 | 67.7% (0.65%) | 31.5% (0.61%) | 68.1% (1.59%) | 43.1% (0.78%) |

|  |  |
| --- | --- |
|  | Features |
| Model 1 | AccountAge, ViewingHoursPerWeek, UserRating, GenrePreference\_Action, GenrePreference\_Comedy, GenrePreference\_Drama, GenrePreference\_Fantasy, GenrePreference\_Sci-Fi, AccountAgeBins\_0, AccountAgeBins\_1, AccountAgeBins\_2, AccountAgeBins\_3, AccountAgeBins\_4, AccountAgeBins\_5, AccountAgeBins\_6, AccountAgeBins\_7, AccountAgeBins\_8, AccountAgeBins\_9, UserRatingBins\_(0.996, 2.0], UserRatingBins\_(2.0, 3.0], UserRatingBins\_(3.0, 4.0], UserRatingBins\_(4.0, 5.0] |
| Model 2 | AccountAge, MonthlyCharges, ViewingHoursPerWeek, AverageViewingDuration, ContentDownloadsPerMonth, SupportTicketsPerMonth, PaymentMethod\_Mailed check, PaperlessBilling\_No, ContentType\_TV Shows, MultiDeviceAccess\_No |
| Model 3 | AccountAge, MonthlyCharges, TotalCharges, ViewingHoursPerWeek, AverageViewingDuration, ContentDownloadsPerMonth, UserRating, SupportTicketsPerMonth, WatchlistSize, SubscriptionType\_Basic, SubscriptionType\_Premium, SubscriptionType\_Standard, PaymentMethod\_Bank transfer, PaymentMethod\_Credit card, PaymentMethod\_Electronic check, PaymentMethod\_Mailed check, PaperlessBilling\_No, PaperlessBilling\_Yes, ContentType\_Both, ContentType\_Movies, ContentType\_TV Shows, MultiDeviceAccess\_No, MultiDeviceAccess\_Yes, DeviceRegistered\_Computer, DeviceRegistered\_Mobile, DeviceRegistered\_TV, DeviceRegistered\_Tablet, GenrePreference\_Action, GenrePreference\_Comedy, GenrePreference\_Drama, GenrePreference\_Fantasy, GenrePreference\_Sci-Fi, Gender\_Female, Gender\_Male, ParentalControl\_No, ParentalControl\_Yes, SubtitlesEnabled\_No, SubtitlesEnabled\_Yes, AccountAgeBins\_0, AccountAgeBins\_1, AccountAgeBins\_2, AccountAgeBins\_3, AccountAgeBins\_4, AccountAgeBins\_5, AccountAgeBins\_6, AccountAgeBins\_7, AccountAgeBins\_8, AccountAgeBins\_9, UserRatingBins\_(0.996, 2.0], UserRatingBins\_(2.0, 3.0], UserRatingBins\_(3.0, 4.0], UserRatingBins\_(4.0, 5.0] |

## Model Selection

Based on the model summary comparison, I selected Model 2. The performance is significantly higher across the board. Interestingly, in previous experiments, I found that model 3 had slightly better performance until I implemented the new undersampling technique. Model 2 has only 10 predictor variables, allowing for simpler and more performant production code.

Furthermore, Model 2 was the only one which was trained on the full dataset, while Model 1 and 3, due to processing speed issues, could only be trained on a random sample of 50,000 data points. This means that Model 2’s F1 value is more reliable than the other ones.

This model could be improved by doing due diligence on the feature selection process. I must admit that my focus was to find a “hack” that would give me considerably higher F1 scores, rather than refining the quality of the features. By finding more relevant features, we could probably take the mean F1 score to around 70-72%, as the model occasionally reaches these F1 values.

## Model Interpretation

Although our final model doesn’t have a simple equation, we can still take a look at the related coefficients for each of the features:

|  |  |
| --- | --- |
| Feature | Coefficient |
| **AccountAge** | **-2.1182** |
| **MonthlyCharges** | **1.2301** |
| **ViewingHoursPerWeek** | **-1.3251** |
| **AverageViewingDuration** | **-1.5460** |
| **ContentDownloadsPerMonth** | **-1.3933** |
| **SupportTicketsPerMonth** | **0.7956** |
| PaymentMethod\_Mailed Check | 0.0589 |
| PaperlessBilling\_No | -0.0065 |
| ContentType\_TV Shows | -0.0216 |
| MultiDeviceAccess\_No | -0.0290 |

Based on the coefficients, we can see that of all the selected features, only 6 had a considerable predictive power. By far, the most important factor is the Account Age, which has a negative coefficient. This means that as the user’s account gets older, it’s less likely the user will churn.

Quite surprisingly, Monthly Charges have a negative impact on the user churning. One might think that due to sunk-cost fallacy, people who pay more would be less likely to churn, but the opposite seems to be the case. On top of that, the predictive strength of this feature is rather high.

There is a negative relationship between churn rates and 3 important usage statistics: Weekly viewing hours, length of each viewing session and monthly content downloads. This means that customers that make use of the service to the fullest, are much less likely to churn.

Finally, there is a fair positive relationship between churning and making more support tickets. This is what we had predicted in our data analysis. People that are having problems with the service, are more likely to churn.

## Summary

The most significant variables that affect whether a customer will churn, are the Account Age, the Monthly Charges, the Viewing Hours per Week, the Average Viewing Duration, the Content Downloads per Month, and the Support Tickets per Month. The other variables have a much smaller impact; however, everything helps to make a stronger predictive model.

We were also able to quantify the impact of each factor, as explained in our model interpretation.