In-Flight Media Content Optimization

California State University, Fullerton
In collaboration with Panasonic and Black Swan Data

Louis Bensard, Roxxanne Hobart, Kevin Mori, Mydoris Soto, Wanyi Dai, Ping Zhao, & Nuno Malta
May 5, 2019

Abstract

Airlines invest in media content to be loaded on their aircraft with the goal of creating passenger experiences that are not only enjoyable, but develop customer loyalty. The media items loaded on the aircraft is called 'media load'. While there are many factors that go into making a flight enjoyable for passengers, the media options available plays a role in their experiences. Our team has created a media load recommendation algorithm that predicts proportion of views for each media item. Through generalized linear modeling and k-fold cross validation using proportions of views as our response variable trained on correlation squared, we obtain models that increase the probability of removing media items that are not desired by passengers and add media items that will have high proportion views on future flights.

1 Introduction

Our goal is to provide a media load recommendation for in-flight entertainment by predicting which items are preferred by passengers. In specific we developped recommendations that suggest items to remove from the media load as well as which to add. These will later be identified as Recommendation I and Recommendation II respectively. Strategically suggesting which items should no longer be invested in will cut overall costs without sacrificing quality in-flight entertainment. We will provide simulations that demonstrate this result.

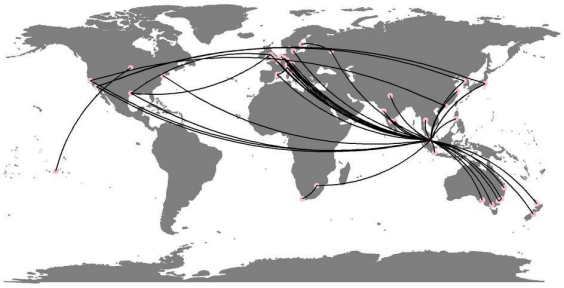

Figure 1: Singapore Airlines Routes (Oct. 2018 - Feb. 2019)

The history data we will analyze is from 5 months of previous flights for Singapore Airlines. These routes from October 2018 to February 2019 are mapped in Figure 1. While the data has been cleaned slightly, Table 1 and Table 2 display a sample of variables which we utilize to predict *proportion views*. Ultimately, after organizing the data, we are able to fit a Generalized Linear Model to the response variable *prop_views* with the provided media variables as explanatory variables. From this prediction we are able to predict the success of new media items that have never been tested on the airline media load. This will offer as a recommendation of items to add. A strategic data analysis of the data history couple with data engineering, we are able to provide a 'remove' list for which media to remove from the current media load on planes. Before we began our analysis, we needed to do some data organizing including: dimension reduction, data manipulation, aggregation and develop methodologies for replacing missing data and augmenting the given dataset.

2 Data Organization

The airline history data we analyzed contained what we separated into two categories: Flight Data (table 1) and Media Data (table 2). Those two tables are linked using *Media ID*, a unique ID identifying each media.

Table 1: Sample of Flight Data

Flight No.	Departure Airport	Seat No.	Departure Date	Media ID	Watch Duration
421	India	48C	2018-11-05 05:32:29	sqm071800014m4	98
337	Germany	15F	2018-12-07 08:26:14	sqs121800067m4	20
236	Australia	31D	2018-12-12 01:01:13	sqm111800009m4	60
51	United States	11A	2019-01-12 22:39:47	sqm011900006m4	111
236	Australia	33H	2018-11-03 03:46:06	sqs111800082m4	52
962	Singapore	55H	2019-01-20 06:09:49	sqs091800051m4	21
32	Singapore	49D	2018-10-14 01:12:37	sqm101800054m4	44
32	Singapore	31E	2018-11-09 00:11:04	sqm071800008m4	97

Table 2: Sample of Media Data

Media ID	Title	Type	Year	Genre	COO	PPL Score
sqm071800014m4	The Full House	movie	2018	Comedy	FRA	5.4
sqs121800067m4	Petey IV (S01, Ep7)	tvepisode	2017	Animation	USA	8.3
sqm111800009m4	Teen Titans Go! To the Movies	movie	2018	Family	USA	6.9
sqm011900006m4	White Boy Rick	movie	2018	Drama	USA	6.5
sqs111800082m4	S01, Ep 1 The Split	tvepisode	2018	Drama	GBR	7.1
sqs091800051m4	One With The Red Sweater (Ep2)	tvepisode	2001	Comedy	USA	9.0
sqm101800054m4	Tarzan (1999)	movie	1999	Family	USA	7.2
sqm071800008m4	Anon	movie	2018	Thriller	DEU	0.0

The flight data includes variables on the flight such as *Flight number*, *Departure (and arrival) airport*, *Departure (and arrival) date*, *Flight duration*, *Seat number* along with media related variables such as *Media ID*, *Watch duration*. Later we discuss the benefits of incorporating the *Seat class* variable (Business or Economy) as an option for airlines to put a weight on the business class as to further assure the higher paying customers are getting the media load they prefer. As we walk you through the data organization we discuss data cleansing. We also discuss the aggregation and identify the response variables for supervised learning.

The original Media Dataset consisted of 1933 movies and TV shows. Most of the basic information about the media (*title*, *realease year* etc.) was here but we will see later on that we had to use data augmentation teechniques in order to have a richer and more useful Media Data.

Our data cleansing centered on three key areas identified as weak points in the dataset:

1. Basic dimension reduction
2. Feature engineering
3. Data augmentation: for both media data and flight data

2.1 Data Cleansing: Basic Dimension Reduction

The original Flight Dataset we received is vast: 10.3 million observations on 41 variables. (Each observation is a passenger watching a media). Some of those variable are not useful for our recommendation system, and some have confusing and repeated verbiage that needed to be fixed. It was a lot of little things to fix and tweak that adds up to an extensive amount of data manipulation and cleansing. A example would be the watch duration of a movie sometimes exceded the movie duration, which is impossible. As another example, four columns for arrival airport information were found (IATA code, ICAO code, Airport Country, and the Airport name). With a data set that large, it was necessary to spend time reducing these. This basic dimension reduction helped reduce the dimensionality of the dataset significantly and made it more usable.

2.2 Data Cleansing: Feature Engineering

We also wanted to find more pertinent information in the data that were not (specifically) listed. The feature engineering will ensure we are able to quantitatively use many of the variables provided to us.

Seat Class

As we mentioned before, there may be a use for identifying seat class. However, the data we received only has seat number. So we took the *Seat number* variable and matched it to a seat location instead by researching the seat locations for a 787-100 Boeing. Our previous list of raw seat numbers has now been replaced by *seat class*: "business" and "economy" as shown in Table 3.

Table 3: Sample of Seat Class

Seat No.	Seat Class
52A	Economy
48C	Economy
15F	Business
31D	Economy
11A	Business
33H	Economy
55H	Economy
49D	Economy
31E	Economy

Proportion Viewed

A metric that will be central in our recommendations will be *proportion views* (or *prop_view*). This variable quantifies the probability that a movie will be clicked

on at all. This proportion is of how many times the media item was clicked on compared to all *clicks* for the entire dataset. For instance, the dataset has a total of 3 million total views, thus if 'The Incredibles' was viewed 30,000 times, then it would have a proportion view of 1%.

However, the dataset covers a total of 5 months and some media might have been on the planes for 5 months while others might be loaded for only one month. We call *cycles* the number of months a media has been loaded for. We need to take those cycles into account so that no media gets advantaged by simply be more cycles on a plane than other media. Therefore, we normalize the *proportion views* mentionned above by dividing that proportion by the number of cycles of the media.

Lastly, an airline might want to put more weight on the satisfaction of the Business Class compared to the Economy class. This is why we added the the weight *business_weight* when computing *prop_view* in the aggregation part. We currently set this weight to 1.5 (meaning the Business class is 50% more important that the Economy class) but it can be modified by the user. The *proportion views* mentionned above is then simply multipied by that weight if the passenger is in Business Class.

As a result, for media 'title' i:

$$prop\_view[i] = \frac{views[i] * \begin{cases} 1.5, \textit{ if Business Class} \\ 1, \textit{ Else} \end{cases}}{cycle[i] * \sum_{i=1}^{1933} views[i]}$$

Figure 2 shows the distribution of *prop_view*. The main takeaway from this plot is:

'A minority of media are watched by a majority of passengers'

Keep this in mind all throughout this paper as it will become central later on in the two stages of our recommendation system. (e.g: most of the area under the curve corresponds than less than 10% of the media load).

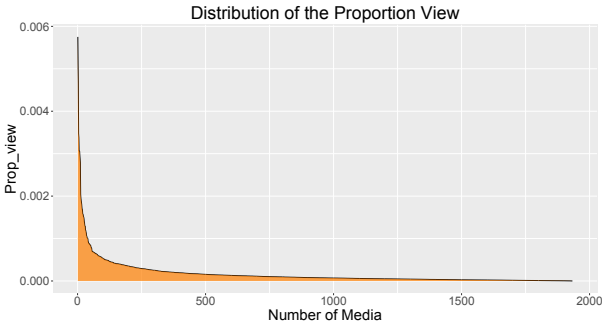

Figure 2: Distribution of Proportion Views

Proportion Used

Another variable we re-coded is proportion of media watched. This is in lieu of, and calculated from, the seconds watched. We created it so that the usage of different movies can be compared. Movies don't have the same duration. A passenger can watch 100% of a 1-hour movie and 50% of a 2-hour movie and both are recorded as 1 hour of used time. We modified this value by dividing it by the full length of the media item. This will be more helpful for determining which media items are enjoyed by passengers. We call this variable *proportion usage* and label it *prop_usage*. For the $i = 1, 2, ...1933$ media titles:

$$prop\_usage[i] = \frac{play\ duration[i]}{full\ media\ duration[i]}$$

For example, if a media item is 6000 seconds, but the passenger only watched 600 seconds of it, that proportion would be 10% or 0.1 because they turned the item off early.

Figure 3 shows the density of *prop_usage* (before aggregation). We can clearly see two humps. The first one corresponds to the 'miscliks', people who clicked on the wrong movie or change their mind after few minutes of watching. The second humps corresponds to the expected usage when watching movies or TV shows.

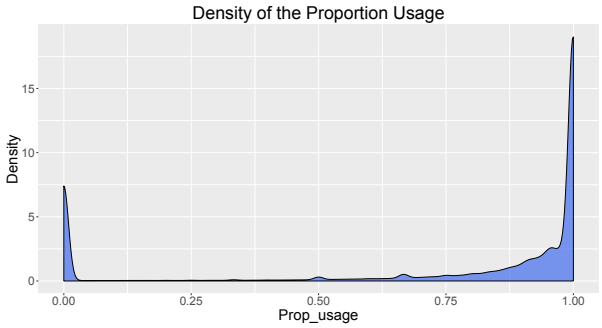

Figure 3: Density of Proportion Usage

Table 4: Sample of A-List Quantity

Title	Cast	A_List_Qty
I Feel Pretty	Amy Schumer, Michelle Williams, Tom Hopper, Rory Scovel, Adrian Martinez	2
American Animals	Evan Peters, Barry Keoghan, Blake Jenner, Jared Abrahamson, Ann Dowd	0
The Seagull	Saoirse Ronan, Corey Stoll, Elisabeth Moss, Annette Bening, Mare Winningham	3
Please Stand By	Dakota Fanning, Toni Collette, Alice Eve, River Alexander, Marla Gibbs	3
Edie	Sheila Hancock, Kevin Guthrie, Paul Brannigan, Amy Manson, Wendy Morgan	0
Taxi 5	Franck Gastambide, Malik Bentalha, Bernard Farcy, Salvatore Esposito, Edouard Montoute	0
Strawberry Days	Nelly Axelsson, Ola Cywka, Stanislaw Cywka, Jan Dravnel, Åsa Håkansson Strömberg	0
Dude's Manual	Dong Zijian, Zhong Chuxi, Jessie Li	0
The Faces of My Gene	Yue Yunpeng, Wu Jing, Wu Xiubo, Jing Boran, Lin Chi-Ling	0

Later we discuss the outcomes of using these two variables and identify *proportion views* as a much more useful variable. Next we describe aggregating the data by media *title* over all the flights. It is during this aggregation that the two variables *prop_view* and *prop_usage* are computed.

Release Year

In addition to the data modifications described above, we also categorized the *release year* of each media into 7 categories. These were 'new release', '1 year old', '2-4 years old', '5-8 years old', '9-14 years old', '15-20 years old', and 'more than 20 years old'.

Price Tag

From the variables *release year category* and *people score* we created a *price tag* variable based on these values. Every media gets its own price tag and it would correspond to how much the airline has spent on that media. These prices are based on judgment since, though they are available, we were not provided with actual price tags. We encourage those that utilize the algorithm to better its validity by using actual prices of the media.

Actors

Our biggest feature engineering challenge was using the actors data. In Table 4 you can see the string of actors in the cast variable. This string is unique for almost every media *title*. With every entry unique it will have no value on predicting *proportion views* and *proportion usage*.

In order to make the actor data useful, we decided to recode this variable into a numeric variable indicating the number of "A-List" stars acting in the title. By identifying quality actors we can better quantify their value.

As a source for this process, we used Box Office Mojo which is a website that tracks box office revenue in a systematic, algorithmic way. The site was founded in 1999, and was bought in 2008 by IMDb, which itself is owned by Amazon. The website is widely used within the film industry as a source of data. But the API is a paid service, so we employed a variety of web scrapers to get the list of actors and directors that grossed over $1 billion (our arbitrary definition of what defines "A-List").

Using the actors field, we extracted a vector of actor names and cleaned the names (trailing whitespace, uppercase, punctuation, etc). Box office mojo has a list of the highest grossing actors and directors. Next, we matched the airline actors data with the top grossing actors on last name (using a hard match) and first name (using a fuzzy match). If there was a match, it's recoded into a 1. If there isn't a match, it's recoded as a 0. For each movie, we sum them to get the number of A-list stars.

Nominated or Won Awards

We also identified from the API which media titles were nominated for (and won) awards using regular expressions to search for keywords "nominated" and "win/won". The variable *nom_idx* and *win_idx*w identifies these respectively and are displayed in Table 5.

Table 5: Sample of Award Winning Media

Title	Nominated	Won
Beast	1	0
American Animals	1	0
The Seagull	0	0
Edie	1	0
Taxi 5	0	0
Spotlight	1	1
Strawberry Days	0	0
Dude's Manual	0	0
The Faces of My Gene	0	0

2.3 Data Cleansing: Data Augmentation

Media Data

Our first batch of data, representing one airplane's flights in the month of November 2018 required a high degree of data cleansing because the some a significant amount of flights had partial missing data and the media data was incomplete.

We didn't want to exclude these titles and potentially bias the results, so for our first batch, we had to use outside sources to augment the dataset. We took a creative approach to fill the missing values in order to retain valuable observations on every media *title*.

Our first external dataset comes from IMDB itself. In lieu of an API, IMDB makes their data available in bulk files (https://www.imdb.com/interfaces/). We were able to use this file to fill in missing *Country*, *Year*, and *People Score* variables by matching against the title.

If we were unable to find a title match on a piece of media, we used the OMDB search API to find titles and augment the country and year variable. Using this methodology, we were able to positively identify 98% of missing variables.

In the second data drop, we were able to acquire synchronized media load and media usage datasets, which rendered the above algorithm obsolete (we were able to get metadata for 100% of the entries using the updated media load data). However, we recommend the user to use the methodology mentionned above if there is a significant amount of missing data.

Airport and Location Data

Using the Open Airports dataset, we were able to translate the ICAO airport code to geographic country, latitude, longitude, and common airport name. We do not use those variables in our analysis but they might be helpful to improve recommendations as mentioned in the Next Steps section.

2.4 Aggregating

In order to quantify how much of each of the media items were watched on all the flights we need to aggregate based on the individual media *titles*. The history data from Singapore Airlines was a list of everything that was ever clicked on for all five months of flight routes shown in Figure 1. The resulting aggregation from our algorithm is a list of the 1933 unique media *titles* and their according *proportion views* and *proportion usage*, for all flights on all five months. Both of these variables are critical: 'proportion views' will help quantify what type of media items passengers are looking for, while *proportion usage* will help quantify how much the media item was enjoyed.

2.5 Supervised Learning

While there is no specifically identified response variable, it is natural to consider the proportion of media watched and the number of times a media item was viewed (clicked on) on as the variable that represents the passenger satisfaction with each media item. For the remainder of the analysis we assume the *proportion views* (*prop_views*) and *proportion usage* (*prop_usage*) represent the satisfaction of the passenger in order to follow through with supervised learning.

The recommendation system we provide is divided into two parts. First, we will suggest what media the airline should immediately unload from the planes and not buy in the future ('Remove' Recommendation). Second, given a list of new Movies and TV Shows and airlines specified goals, we suggest a subset of that list that the airline should invest in and load onto the specific aircraft ('Add' Recommendation).

We are only interested in predicting proportion views and proportion usage for movies and TV shows, so we disregarded the other media types. Further analysis could potentially include what percentage of media items are watched compared to games and music as to quantify the need for quality movies and TV Episodes.

Let's take a closer look at how *proportion views* and *proportion usage* can be useful in order to provide a sound recommendation. Imagine Movie 1, a movie that was only watched 200 times out of 3 million but fully watched (100% usage). Now picture Movie 2, a movie that was watched 200,000 times out of 3

million and but people on average stopped at 70% through the movie (70% usage). What movie would you recommend? Movie 2 is clearly more valuable than movie one for an airline company.

This is why we decided our creteria for a 'good media' will be determined principaly by *proportion views*. It is only when we decide that a movie has a high enough *prop_view* that 'proportion usage' can come into play. See Figure 4.

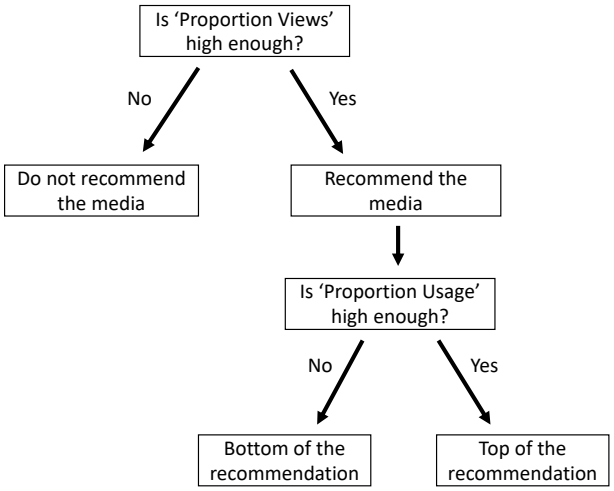

Figure 4: Hierarchy of proportion views and proportion usage

We will later expand on what makes *proportion views* or *proportion usage* 'high enough'.

3 Optimization and Prediction

Recall the first part of our recommendation system. Givin a list of unseen and unloaded media content, the task at hand is to optimize the choice of those media so that they can be loaded on an aircrafts. While we do want to *optimize*, we first need to *predict* the outcomes of our response variables (y_i's) *proportion views* and *proportion usage*.

Indeed, those media have never been loaded on planes before and thus we do not know those two proportions. With the data organized as described above, we are now be able to predict these response variables for each media item. The method we use for optimization and prediction is generalized linear models with k-fold cross validation.

3.1 GLM: The Theory

To explain GLM, it is easiest to first draw out the most basic form of classic linear models. Here we have the equation representing the relationship between the y_i's (response variables) and x_{ij}'s (predictor variables) for $i = 1, ...n$ samples and $j = 1, ...p$ predictor variables in a random sample as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{12} + ... + \beta_p x_{ip} + \varepsilon_i \qquad (1)$$

Where the ε_i's are errors and y_i's are linear functions of the x_{ij}'s. We can estimate the coefficients β's using least squares regression.

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (2)$$

In classic linear models we assume the y_i's are normally distributed random variables with constant variance and $\mathbb{E}(Y) = \mu = X\beta$. The model fit can be quantified by the RMSE (variance of model errors) and r^2 values (variance explained in the response variable by the predictor variables).

Generalized linear models are like linear models except they allow for non-normal errors in the response variable as well as allow us to account for different types of distributions on y_i, which is often (and in our data) the case. They do this by allowing the linear model to be related to the response variable via a link function, and by allowing the variance of each measurement to be a function of its predicted value. The generalized linear model equation is:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \qquad (3)$$

Where g is the link function and the mean of $y_i|\mathbf{x}_i$ is μ_i (or $E(y_i|\mathbf{x}_i) = \mu_i$). The x_{ij}'s are the $j = 1, 2, ...p$ predictor variables for i observations such as genre, release year, critic score, etc.

Binomial with logit link function

Since our response variable is a proportion (between 0 and 1), we use family=binomial(link=logit). This can be visuilized in Figures 2 and 3. Here we briefly describe logistic regression for generalized linear models.

From Equation (1) and Equation (3), we are assuming that a transformation of the conditional expectation $E(Y|X)$ is a linear function of X

$$g(E(Y|X)) = \beta^T X$$

for some function g. We can see that the transformation was the identity transformation $g(u) = u$ in linear regression. Now for logistic regression we use the logit transformation $g(u) = log(u/(1-u))$.

Given predictors $X \in R^p$ and an outcome Y, GLM is defined by three components: a random component $(Y|X = \mu)$, a systematic component (β), and a link function (g) that connects the random and systematic components. Let $\eta = g(\mu)$ from equation (3). The systematic component relates a parameter η to the predictors X. Then, a generalized linear model is given by

$$\eta = \beta^T X = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

In particular, the link function provides a connection between μ and η. For example, the normal case, where $g(\mu) = \mu$, so that $\mu = \beta^T X$.

However, for our data we are not using the identity link function g. For GLM, we are always modeling a transformation of the mean by a linear function of X. For the Bernoulli case, the "right" choice of link function is logit transform

$$g(\mu) = log(\frac{\mu}{1-\mu})$$

Therefore, we have the model

$$log(\frac{\mu}{1-\mu}) = \beta X$$

Our data is a set of independent data $(\mathbf{x}_i, y_i)$ for $i = 1, 2 \cdots n$. Our goal is to estimate $\mu_i = E(y_i|\mathbf{x}_i)$. Recall that we have a link function $g(\mu_i) = \eta_i$, which connects the mean μ_i to the parameter $\eta_i = \mathbf{x}_i^T \beta$. We can first estimate the coefficients β using $\hat{\beta}$ (see equation (2)):

$$g(\hat{\mu}_i) = \mathbf{x}_i^T \hat{\beta} \quad for \quad i = 1, \cdots n \tag{4}$$

That is

$$\hat{\mu}_i = g^{-1}(\mathbf{x}_i^T \hat{\beta}) \quad for \quad i = 1, \cdots, n$$

The glm() function (in R or in Python) uses maximum likelihood to compute $\hat{\beta}$. With generalized linear models we can easily get these coefficient estimates for each media variable and model the proportion of media viewed and the proportion of usage. In our code on this method, we use the Binomial family and a the logit link function as we just discussed.

3.2 GLM: The Application

Our algorithm is using glm() in order to predict the value *proportion views* ($y = prop\_view$) for a new list of media. The characteristics of the media are the explanatory variables x_i's of the model in equation (4): x_1=*a_list_qty*, x_2=*win_idx*, x_3=*nom_idx*, x_4=*genre*, x_5=*country_zone*, x_6=*year_category*, x_7=*price_tag*, x_8=*peopleScore*, x_9=*ratingDes*, x_{10}=*contenttype*.

(*ratingDes* is the rating of the media such as 'PG-13' of 'R' and *contenttype* is the type of the media such as 'Movie' or 'TV Episode').

We looked at further dimension reduction and did not get better results from removing any of the predictor variables listed above, and so we kept all of them in the GLM.

Our response variable $y = prop\_view$ is between 0 and 1. This constraint led us to choose the binomial(link=logit) distribution for the response in the glm(). More complex generalized linear models using the Beta distribution were unsuccessful and performed poorly compared to the binomial.

3.3 GLM: Available Packages

There are multiple handy packages in R and Python to implement general linear regression models.

In R, the popular MASS package can be applied for the GLM regression model by calling the glm function. When using the glm function, the main arguments that need to be specific including Formula, Family, Data.

The Formula is symbolic description of the model to be fitted, it needs to be of the form: response variable ~ feature1+feature2+feature3. The Family is the link function to be used in the model. Some of the possible options are Gaussian, Binominal, or Poisson. The Data is the dataset used to train the model

In Python, "Scikit Learn" is a widely used module for GLM. General steps to train a linear model are following:

1. Import sklearn module, eg: from sklearn import linear_model;
2. Create the design matrix (X) and the response variable (y)
3. Specify the model that needs to be trained, eg: regr = linear_model.LinearRegression()
4. Train the model, e.g: regr.fit(X, y)

5. Make prediction with the new model: regr.predict(X)

If considering parallel computing, H2o is a package available on both R and Python for GLM models.

3.4 k-fold Cross-Validation

The cross validation method is a type of re-sampling procedure that is used to evaluate the model that was fit onto the data. Particularly, it is more focused on how the model is expected to perform when making predictions on the data that is not used during the training step of the model. This type of evaluation method is used when there is limited data available. The algorithm for this method is quite simple and illustrated in Figure 5.

According to many articles and textbooks, there is no formal rule on how to choose the best k. However, there is a bias-variance trade off that is associated with the choice of k. Typically, given the bias-variance trade off, the standard value of k for the cross-validation, is to use $k = 5$ or $k = 10$. We choose $k = 10$ in our recommendation algorithm.

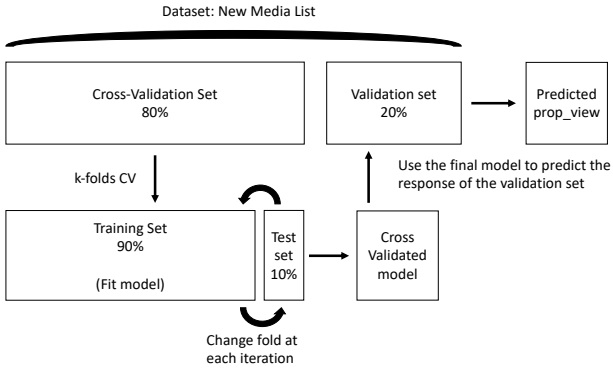

Figure 5: k-fold Cross-Validation

3.5 Model Selection

We first tried different metrics such as r^2, BIC or ρ^2 to measure model fit (ρ^2 being the correlation squared between the true response y and the predicted response $\hat{y}$).

Ultimately we want to make a recommendation list that recommends removing the bottom $b\%$ of the current media load and add the top $t\%$ from a new list media. So we tried just that; trained our model on accurately classifying the top $t\%$ of the list.

For simplicity lets assume for now we want to add the top $t = 10\%$ of a new list. With each iteration in the k-folds we calculate the number of media items that are accurately identified as being in the top 10%. More specifically, we ordered the data and predictions in decreasing order according to their *proportion views*. Using the response variable y_i : *prop_views* we identify the top 10% and then see how many of the predictions $\hat{y}_i$ from the model were also in the top 10%. For the 1933 media items in the media load there are $0.1 \times 1933 \approx 193$ of them in the top 10%. Let *p_view* be the prediction accuracy that we use to measure model fit when modeling the *prop_view*. For the top $t\%$, its general formula is:

$$p\_view_t = \frac{\sum_i^{1933*t} 1_{y_i \leq \hat{y}_i}}{1933 * t} \tag{5}$$

However, Figure 6 shows quite clearly that all those model selection criteria give close results. Therefore we decided to use ρ^2 as it is more consistent and computationally much faster. Nevertheless, the metric *p_view* introduced above is going to be a key component in providing the final recommendations.

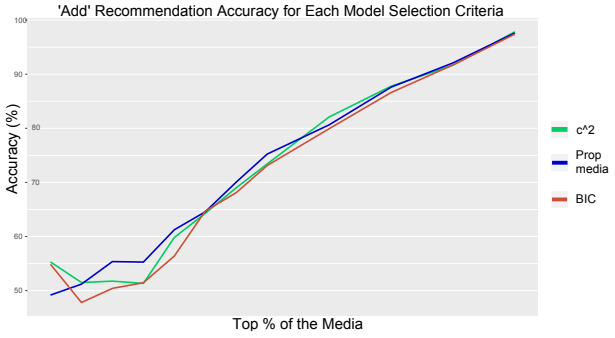

Figure 6: Correlation squared (green), Proportion of media (blue), and BIC (red)

4 Results: The Recommendations

4.1 Recommendation I: Removing Low Performing Media

As previously mentioned in Figure 2, a minority of media are watched by a majority of passengers. The data we have is coming from Singapore Airlines but that statement will be true across all airlines. It is actually a law observed across all industries and known as Pareto's Law or *The 80/20 Principle*. Therefore, using the *proportion views* as our metric for success, it

is possible for us to weed out most of the 'unsuccessful' media without damaging the customer satisfaction. This improvement in the media load would therefore cut media expenses significantly and improve the media selection for passengers.

The question becomes: how much media should we remove? Too few media removed and the airline company would continue wasting money. Too many media removed and the passengers would have less 'successful' media to watch and thus less satisfaction. There is a *Cost-Satisfaction* tradeoff that we need to solve. The problem now becomes finding the sweet spot of proportion of media to remove.

Figure 7 illustrastes what percentage of media explains total views across all flights in the whole dataset. For instance 25% of the media explain about 75% of the usage.

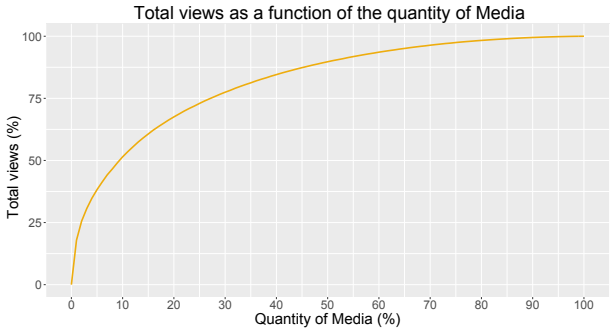

Figure 7: Evolution of the total views as a function of the quantity of media

With this in mind, we are now going to recommend 3 options for the airline company: The *Aggressive* option, the *Moderate* option and the *Conservative* option. Those recommendations are summarized in table 6.

Table 6: Options for Media Removal

Type of Option	Quantity of Media to remove	Remaining Media Usage
Aggressive	80%	68%
Moderate	35%	95%
Conservative	15%	99%

Those 3 options can be changed to the preference of the airline company but should be the first one to consider.

Example

Suppose the airline listened to our *Moderate* recommendation. We now need to remove 35% of the medias. To do so, we simply sort the media by *proportion views* in decreasing order and drop the bottom 35% of the list. That way, we will have dropped the most unsuccessful 35% of the media.

It is important to note that we observed above a large proportion of media with very low *proportion views*. Only in this condition it is acceptable to remove a big part of the media load. Exploratory data analysis before any implementation of our recommendations is necessary.

In this case, we now reduced the number of media from 1933 to 1256 while impacting at most only 5% of the total views.

4.2 Recommendation II: Predicting High Performing Media

The value of our GLM results comes more from predicting new media items to load on aircrafts. Once the final model is trained on the 5 months of data, it will be able to predict the values of *proportion views* and *proportion of usage* for a a brand new set of media that has never been loaded on planes before.

Not knowing what situation the airline company might face, we developped 3 possible scenarios that might occur when wanting to add new media to the existing media load on planes.

The 'validation' step of the cross-validation process allow us to provide a measure of accuracy for the predictions of the model. The way we get this measure of accuracy will depend on each scenario and we will therefore explain it for each of those 3 cases.

4.2.1 Scenario 1: Quantity

The airline company has a list of media they might want to buy and load on airplanes, but they don't want the full list, they want only the top $t\%$ media of that list.

Our model has been trained on 5 months worth of data and thus will be able to give us prediction of the *proportion views* for this new list of media. Once we get the predicted values of *prop_view*, we simply sort the new list of media by *prop_view* in decreasing order and recommend the top $t\%$ of that sorted list.

However, those predictions are not 100% accurate. The more selective the airline company is going to be with regards to the quality of the media (i.e. choice of

t), the lower the accuracy of the predictions. Figure 7 shows how accurate our algorithm is compared to the baseline of randomly selecting media from that list.

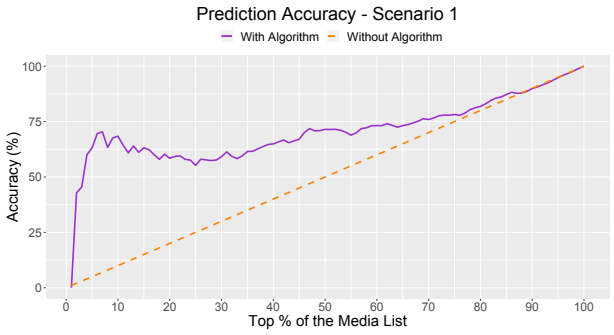

Figure 8: Predictions Accuracy for Scenario 1

This measure of accuracy for our algorithm is computed using the p_view_t value mentioned Section 3.5. This value measures the proportion of media predicted to be in the top $t\%$ that actually are in the top $t\%$ of the list. (We use the validation dataset and thus we know the real value of *prop_view*).

Example

Now, suppose the airline company obtains a new list of 200 Movies and TV Shows. After looking at Figure 8 they decided that they want to buy the top 20% of that list (40 media). See Figure 9 for our recommendation.

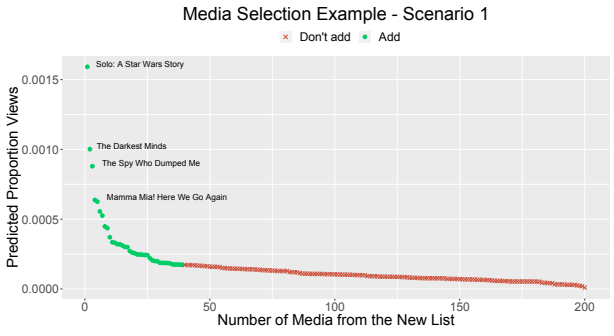

Figure 9: Movie Selection Recommendation for Scenario 1

Figure 10 gives a visual representation of the improvement provided by our algorithm.

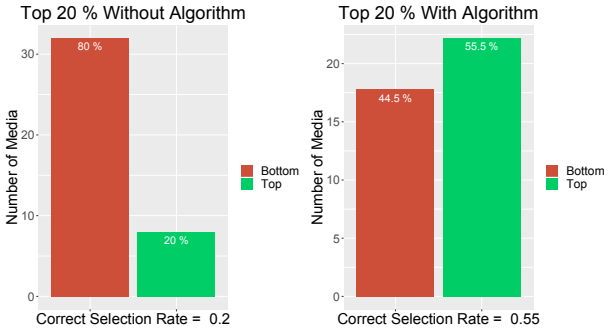

Figure 10: Accuracy comparaison with and without the recommendation for Scenario 1

4.2.2 Scenario 2: Quality

As you might already have noticed, Scenario 1 has a flaw, it ranks media from a new list relative to that new list. What if the list of new media is full of low quality movies? Scenario 1 would still recommend $t\%$ of that list, even thought the media are bad and unwanted by passengers. We need a scenario more flexible that can select media of quality relative to the media usage history. Indeed, while keeping the media load list fresh is good, we need to make sure the list does not get worse in performance by restricting to just the chosen percent.

Here is where Scenario 2 comes into play. We still presume that the airline company has a list of new media and wants to load a subset of them on the aircrafts. However, this time we suppose that the airline company wants to buy all the media that would fall within best $q\%$ of current media load. That is different than just selecting the top $t\%$ of the list. Here we want to select only movies of proved quality from the passenger, hence top *'q'%*.

For instance, say the airline wants the media that would fall within best 20% of current media load. From the current media load data it turns out the top 20% corresponds to a prop_view of 0.0002. Therefore, after predicting all the values of *prop_view* of the media from the new list our algorithm will select all the media that have a prediction *prop_view* greater than 0.0002.

As a result, depending on the quality of the new list, the algorithm might select more or less than $q\%$ of that list. In our opinion, this Scenario is superior than Scenario 1 and should be used when possible by airline companies.

Once again, our algorithm is not 100% accurate and

some misclassification can occur. Figure 11 shows how accurate our algorithm is compared to the baseline of randomly selecting media from that list. It is to be noted that the baseline dotted line is what would happen in a best cases scenario. Indeed, without a strategic data analysis there is no chance of knowing what the top $q\%$ might be.

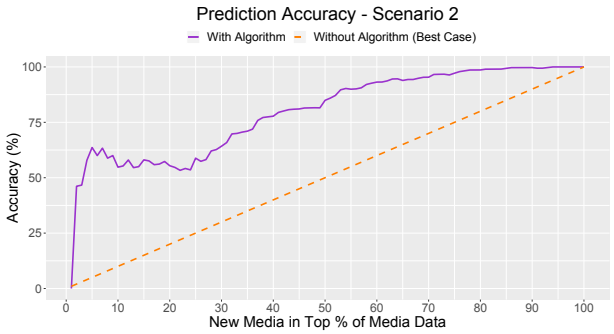

Figure 11: Predictions Accuracy for Scenario 2

This measure of accuracy for our algorithm is very similar to Scenario 1. This value measures the proportion of media predicted to be in the top $q\%$ of the dataset that actually are in the top $q\%$ of the current dataset. (We use the validation dataset and thus we know the real value of *prop_view*).

Example

Let's extend the example mentioned above, where the airline company wants the media that would fall within the best 20% of current media load. The algorithm for scenario 2 recommends 29 media, see Figure 12 for our recommendation.

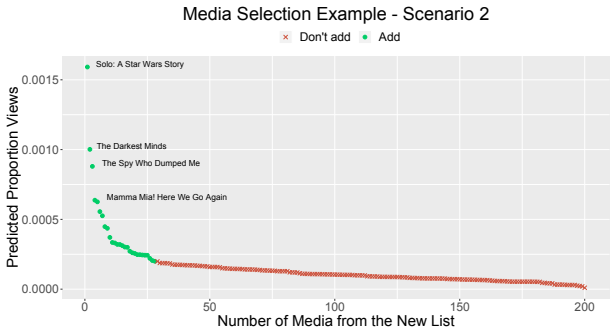

Figure 12: Movie Selection Recommendation for Scenario 2

Figure 13 gives a visual representation of the improvement provided by our algorithm.

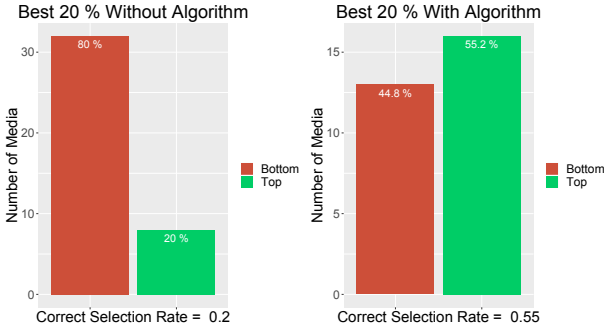

Figure 13: Accuracy comparaison with and without the recommendation for Scenario 2

Note that as expected, the recommendations for Scenario 1 (40 media) and Scenario 2 (29 media) are different. Scerario 1 asked for the top 20% of the new list while Scenario 2 asked for the media in the top 20% of the data history.

4.2.3 Scenario 3: Budget

Now assume that the airline wants to add all the top performing media within a designated budget of *$m*. While we do not have access to the actual cost of each media to the airline company, we use the *price tag* variable we created. Each media has a price of 1,2,3 or 4. The results presented in this money index currency can easily be extended to US dollars once the airline provides the real cost associated with each media.

Figure 14 shows the prediction accuracy of our model as a function of the money spent, compared to the baseline.

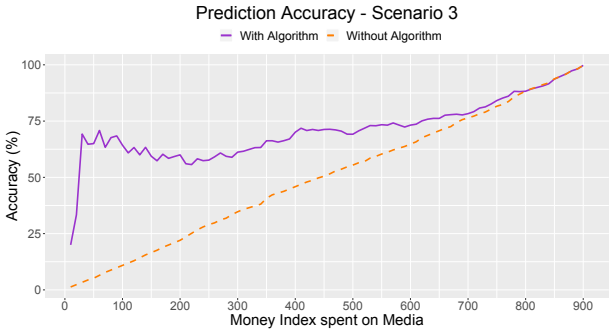

Figure 14: Predictions Accuracy for Scenario 3

This measure of accuracy for our Algorithm is computed the same way as for Scenario 1. However we

first need to translate the amount of money *$m* into a top *t%*.

Example

Using the same new list of 200 Movies and TV Shows, and after looking at Figure 14 with a specific budget in mind, the airline decided that they want to buy $300 (index money) worth of media. See Figure 15 for our recommendation.

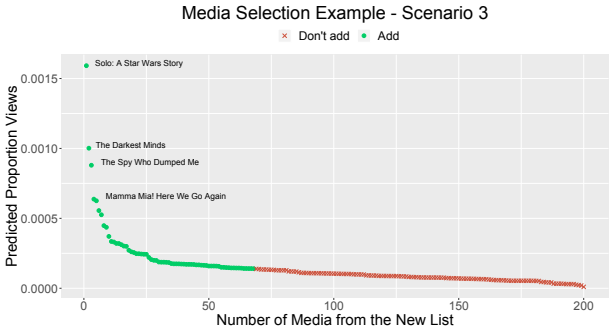

Figure 15: Movie Selection Recommendation for Scenario 3

Figure 16 gives a visual representation of the improvement provided by our algorithm.

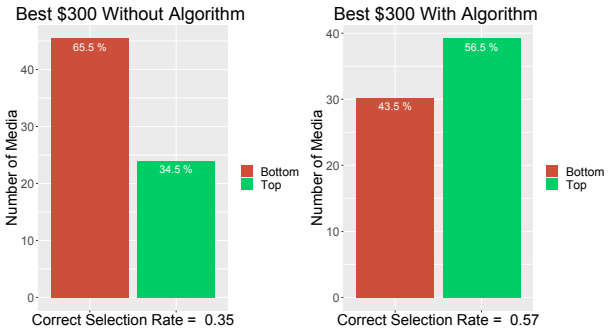

Figure 16: Accuracy comparaison with and without the recommendation for Scenario 3

It is fair to assume airlines don't just guess and have already in place ad-hoc ways of determining which items passengers will watch. However, the results obtained from those three basic scenatio show that there is a lot of room for improvement in the airlines strategy for selecting media to load on planes. Those three senarios aim to give a basic framework for the airline to re-think their strategy of investment in media. They can lead to other scenarios, such as sselecting the media that would be in the top *q%* with a budget of *$m*. That would be a mix of scenario 2 and 3 for instance.

5 Next Steps

Recommendation I: Media Removal

Even though our recommendation for reoming low performing media is very strategic, it is still quite rudimentary and could be improved. Indeed, one the limitation that we have is that we do not know how much each passenger is worth for the airline company. The ideal case for the airline would be to perform a thorough cost-benefit analysis and be able to put a money value for each *view* from a passenger.

The goal would be to put a dollar sign on the satisfaction of a customer. How much money does a happy passenger bring to the company as opposed to an unsatisfied customer? This could be done by analyzing customer loyalty, turnovers, frequent flyers etc. We would advise any airline company to not only keep track of each passenger's ticket purchases but also keep track of each passenger's media usage and be able to link those two together. Surveying their experiences after a flight can be an additional option as well.

Once the value of a satisfied customer has been established, we could estimate the value of a quality media based on the satisfaction of a customer; ultimately creating a variable that quantifies the dollar value of *prop_view*. Maybe 'A Star is Born' might bring back $2 per view to the airline while 'Fat Buddies' might only bring back $0.1 per view to the company. This analysis combined with the associated cost of each media loaded on planes would give us a way to quantify when a media can 'pay for itself' and at least know how much each media is saving money for the airline overall.

Recommendation II: Add New Media

The kind of cost-benefit and return on inverstment (ROI) analysis can also be very valuable to select new movies that we want to load on the planes. Indeed, the best case scenario would be to make *Recommendation I* obsolete by only adding quality movies that would maximize the ROI.

Using our model to predict the *proportion views* of the new media, we could perform the same analysis mentioned above and determine a dollar value per view for each new media. We could then predict the ROI of each new media and select media that meet a minumim threshold of ROI.

Load media depending on Flight Route

Another limitation that we faced is that even though we had access to the flight route a media was watched on, we could not use this information to to give recommendations on a route level. Indeed, even if we found that people traveling from Singapore to Australia loved comedy movies and people traveling from Japan to the US loved drama movies, the airlines would not be able to use this recommendation (at least not all airlines). We belive that the airlines should update their hardware and software in order to make possible the load of different media for each different route. This would make the accuracy and relevance of media selection and media removal recommendations increase significantly.

Similarly, if we get multiple years of data and aggregate on each month we can find correlations between media preferences for each season. We could potentially identify features from November flights 2018 and November flights 2019 that are due to seasonal reasons.

Even more advanced, we also have access to what day of the week the media are watched, business travelers often fly on weekdays and leisure traveler often fly on weekends. Making available the right media to the right people (business travelers or leaisure travelers) by turning on or off specific media during the week would also improve the customers satisfaction by accomidating for their preferences.

Proportion Usage

As discussed we did not follow through with the variable *proportion usage*. This is a response variable worth looking into more. It would be interesting to combine the two variables (*prop_view* and *prop_usage*) as to assure the preferred media are both identified as viewed and watched through completely.

Data Augmentation

Augmenting the media data to obtain variables such as 'A-list actors' or IMDB scores was a immense improvement. Keeping up in that direction and coming up with additional ways of categorizing a media will increase the predictions accuracy. For instance, a variable that can quantify/classify A-list Directors would be valuable.

Conclusion

We started the project and analysis with the goal of providing a media load recommendation that will help airlines select a subset of their current media load as well as recommend new items to add. While our algorithm is not 100% accurate at predicting *proportion views*, it is vastly better than the current methodologies for selecting media and displayed in Figures 10, 13, and 16. As we worked our way through deciding how to best identify which media to remove and add, we tried to put ourselves in the shoes of an airline company. *Recommendation I* offers flexability in the removal recommendation with guided strategy to make the best choice for their company. The three scenarios of *Recommendation II* allow an airline to customize the algorithm to their individual needs, visions, and budgets. No matter the recommendation chosen, we are confident the algorithm will both curate the media load to contain higher percentage of prefered items as well as reduce investment costs for the airlines.

References

Carnegie Mellon University, Department of Statistics and Data Science. Generalized Linear Models (Spring 2014).

James, Gareth, et al. An Introduction to Statistical Learning: with Applications in R., pg 181-191. Springer, 2017

Joseph C. Gardiner, Zhehui Luo and Lee Anne Roman. Fixed effects, random effects and GEE: What are the differences? (2009) *Statist. Med.* , 28:221-239

Russell, Stuart J., and Peter Norvig. Artificial Intelligence a Modern Approach. Pearson, 2018.

https://en.wikipedia.org/wiki/Pareto_distribution

https://scikit-learn.org/stable/modules/linear_model.html

https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/glm