Journal of Information Technology Management

ISSN #1042-1319

A Publication of the Association of Management

PREDICTING AUDIENCE COMPOSITION FOR NEW RELEASE MOVIES WITH LOYALTY PROGRAM DATA

KATHERINE GOFF INGLIS
RYERSON UNIVERSITY
kgoff@ryerson.ca

SAEED ZOLFAGHARI
RYERSON UNIVERSITY
zolfaghari@ryerson.ca

ABSTRACT

Predicting the performance of new release movies with the absence of sales data is a challenging and interesting problem. Intuitively and empirically it is known that heterogeneous groups of movie-goers with unique preferences and tastes exist. In this study, we use loyalty program data to predict audience demographic composition for new release movies using the demographics and past purchase behavior of loyalty card holders. We develop 11 regression models to predict audience composition and identify movie preferences for key demographic groups. Our results align with the existing studies based on customer survey data and provide directional insights to demographic appeal and audience composition for new release movies. Our method using customer loyalty data can be utilized to develop intelligent systems for new entertainment products that are embedded to recommend and inform strategic decisions around affiliate advertising, target marketing and operations management.

Keywords: movie exhibition, new product forecasting, loyalty program, regression analysis, customer marketing, demographic targeting, customer behavior, entertainment products

INTRODUCTION

Forecasting demand for new release movies is a topic that has been widely studied in academic literature. There are a number of studies (e.g. Litman [5]; Ravid [8]; Sharda & Delen [10]; Smith & Smith [12]; Wallace & Holbrook [14]; Zufryden [16]) that quantify the impact of various factors, such as star power, movie-critic reviews, genre, ratings and seasonality on North American box office performance. Predicting demand at a finer level of detail, by movie theatre or customer segment, is an emerging niche within the broader topic of movie forecasting. Somlo et al. [13] use Census data to demonstrate how demand by movie theatre varies as a function of the local population's characteristics within the trade area. The study by Redondo & Holbrook [9] uses customer survey data to show that heterogeneous moviegoer groups with varying demands for specific movie features exist. This study extends the work of Redondo & Holbrook [9] in a new and interesting way exploring the definition of the characteristics of heterogeneous moviegoer groups using customer loyalty data. Through this approach, conventional wisdom is reaffirmed and new insight is gained into the relationship between the key film attributes (genres, movie ratings, critic reviews, production budgets, film length) and sub-segments of the movie-going population.

There are several practical use cases for predictive audience insights to inform the decision making processes of industry stakeholders (such as producers, distributors exhibitors, customers) (see Somlo et al. [13] for a full explanation of the functions of the different industry stakeholders). Many stakeholders obtain this information from a media industry ratings and measurement leader, Nielsen, which offers pre-release customer research in major American markets to compile a number of metrics for new release films, such as interest for key demographic and ethnic segments. Movie industry stakeholders utilize this information to inform key operational, advertising and marketing processes for new release films. The key audience for new release films can be utilized to make scheduling decisions like whether the film should play at the theatre, how many shows should be scheduled at that location or what size auditorium should the film play at that location. Media executives can use the key audience by movie to place relevant advertising with films based on the advertiser's target customer profile. Moreover, the predicted composition of the audience can be used in conjunction with demand forecast data to provide reach and frequency predictions by film for granular demographic segments. The target marketing applications for the film itself are vast. For example, demographic customer segments are used to inform decisions around:

- affiliate advertising with complementary products targeted at key demographics
- targeted advertising buys on social platforms like Facebook
- targeted mass media advertising, based on key demographics
- personalized digital communications via the loyalty program communication channels (e.g. email, SMS, website, mobile app)

The next sections of this paper describe the parsimonious method used for empirical testing with customer loyalty data, beginning with a description of the relevant academic studies focused on predicting customer segments, followed by an explanation of the dataset, data mining problem and quantitative method used. Following the methodology section, the results are presented and the paper concludes with a discussion of further research opportunities.

RELEVANT LITERATURE

Within the movie industry and academia, the importance and existence of heterogonous segments of movie-goers has been established. Within the industry, Rentrak, Nielsen, Baseline Intelligence and Cinemascore are recognized as data solutions that providers use to gain insight on films and audiences prior to theatrical release. Few academic studies exist within the niche area of forecasting demographics for new movies, the most relevant of which are discussed in this section. The study by Redondo & Holbrook [9] offers a review of a number of other studies to demonstrate the importance and relevance of academic studies focused on moviegoer segmentation.

Much of the quantitative academic literature on the movie industry focuses on macro forecasting of North American box office revenue as a function of movie feature appeal. The study by Somlo et al [13] also demonstrates the importance of factoring in audience demographics when forecasting demand for new release movies at a micro level of detail (as in disaggregate forecasts for individual theatres versus the more often studied aggregate North American level of detail). The authors deem theatre level forecasting "extremely challenging" compared to macro forecasting due to variance in location specific characteristics, evident by examining per capita revenue for the same movie by individual theatre location. Demographics are one type of theatre specific attributes that Somlo et al [13] utilize in their forecasting model, which is ultimately an extension of the parsimonious forecasting model developed by Sawhney & Eliasberg [11]. Somlo et al [13] demonstrate that the new function that estimates model parameters by theatre by movie using theatre specific characteristics, such as demographics, has the potential to optimize distribution planning for theatrical movie releases.

Redondo & Holbrook [9] use customer survey data to empirically show that heterogeneous moviegoer groups with varying demands for specific movie features exist. Their study utilizes canonical correlation analysis to model the relationship between movie features and demographics. The dataset used was derived from movie survey and demographic data collected in Spain. A number of movie features were included: country of origin, genre, objectionable content, stars, promotional effort, critic ratings, as well as demographic variables: gender, age cohort, children present, education, social class, and size of municipality. The results of their study show that the included movie features explain 44.6% of variance in demographic composition.

The exploration of the presence of violence and objectionable content in popular media is the subject of the book *Hollywood vs. America: Popular Culture and the War on Traditional Values* written by film critic Michael Medved [7]. De Vany & Walls [2] explore the same topic empirically in their work titled "Are There too Many R-Rated Movies?" and their results suggest that movie studios produce a large number or R-rated movies, despite G and PG movies being more profitable. De

Vany & Walls' [2] results affirm findings from the study by Ravid [8], which concludes that G and PG ratings are positively correlated with film success. De Vany & Walls [2] and Medved [7] suggest that studio rationale for making so many R-rated movies is driven by an artistic endeavor to earn insider respect and praise in the movie industry. Our study aims to provide further insight into this R-rating conundrum by determining if films for mature audiences appeal to certain demographic segments.

A number of studies have demonstrated that positive critics' reviews are related to movie performance (e.g. Eliashberg and Shugan [4], Basuroy et al. [1]). A positive correlation with film performance and production budget has been demonstrated empirically (e.g. Ravid [8], Litman [5]). Seasonality and correlation to box office performance has also been the subject of a number of empirical studies (e.g. Einav [3]). Our study explores whether these film attributes appeal to certain age or gender demographics.

DATA & METHOD: USING CUSTOMER BEHAVIOR DATA TO PREDICT DEMOGRAPHICS FOR NEW MOVIES

The analysis dataset is created by combining a sample of data from three sources: 1) a film attribute dataset from a movie industry data solutions provider, Baseline Intelligence, 2) a production budget dataset sourced from the-numbers.com (http://www.the-numbers.com/movie/budgets/all) and 3) a transactional customer loyalty dataset from a major movie exhibitor. The dataset includes 350 films released in Canada between April 29, 2011 and April 25, 2014. The release date utilized represents the wide release of the film in North America. The loyalty data is limited to the film's first week of release because the intent of this study is to demonstrate a method that can be used when no sales data is available for a film. After the first week of release, sales data exists and can be used to inform the appeal of a film to segments of moviegoers. Each film in the dataset has at least 5,000 loyalty cardholders attending it during the first week. This rule ensures that small and limited release films are not included in the dataset. Loyalty cardholders represent approximately 30% of the box office ticket sales for the major exhibitor who provided the data; the authors believe that directional insights based on their behavior can be used as a proxy for all moviegoers. The titles in the dataset cover provide a representative sample of movies across seasonality, ratings, release strength and genre. The dependent variables used in model testing and development are percentages between 0 and 1 representing the proportion of the audience that each demographic group makes up. The gender and age of the loyalty cardholder is used to determine categorize behavior with the exception of one; the group titled "Prop_Parents" represents the behavior of children and the adults that take them to the movies. The itemized list below details all dependent variables used:

1. **Prop_Parents** represents the percent of loyalty member cardholders attending the film with children.
2. **Prop_M14_18** represents the percent of loyalty member cardholders attending the film that are males between 14 and 18 years old. Coincides with the age of secondary or high school students.
3. **Prop_F14_18** represents the percent of loyalty member cardholders attending the film that are females between 14 and 18 years old. Coincides with the age of secondary or high school students.
4. **Prop_M19_34** represents the percent of loyalty member cardholders attending the film that are males between 19 and 34 years old.
5. **Prop_F19_34** represents the percent of loyalty member cardholders attending the film that are females between 19 and 34 years old.
6. **Prop_M35_49** represents the percent of loyalty member cardholders attending the film that are males between 35 and 49 years old.
7. **Prop_F35_49** represents the percent of loyalty member cardholders attending the film that are females between 35 and 49 years old.
8. **Prop_M50_64** represents the percent of loyalty member cardholders attending the film that are males between 50 and 64 years old.
9. **Prop_F50_64** represents the percent of loyalty member cardholders attending the film that are females between 50 and 64 years old.
10. **Prop_M65P** represents the percent of loyalty member cardholders attending the film that are males aged 65 and up. Coincides with retirement age.
11. **Prop_F65P** represents the percent of loyalty member cardholders attending the film that are females aged 65 and up. Coincides with retirement age.

Independent variables are selected based on access to data and the academic literature and include: genre (one movie can have more than one genre), MPAA rating, Film Runtime, Production Budget, Seasonality, Big Star Count and Award Nominations. Seasonal and Ratings variables have been adapted from what was found

in the literature to the Canadian market to align with the study dataset. The itemized list below details all independent variables used:

1. **Genre_Action** Binary variable equal to 1 if film genre is tagged with action
2. **Genre_Adaptation** Binary variable equal to 1 if film genre is tagged with adaptation, refers to some type if literature converted to a screenplay
3. **Genre_Adventure** Binary variable equal to 1 if film genre is tagged with adventure
4. **Genre_Comedy** Binary variable equal to 1 if film genre is tagged with comedy
5. **Genre_Drama** Binary variable equal to 1 if film genre is tagged with drama
6. **Genre_Family** Binary variable equal to 1 if film genre is tagged with family
7. **Genre_Horror** Binary variable equal to 1 if film genre is tagged with horror
8. **Genre_Period** Binary variable equal to 1 if film genre is tagged with period, refers to films set in a period of the past
9. **Genre_RomCom** Binary variable equal to 1 if film genre is tagged with romantic comedy
10. **Genre_Romance** Binary variable equal to 1 if film genre is tagged with romance
11. **Genre_Sci_Fi** Binary variable equal to 1 if film genre is tagged with science fiction
12. **Genre_Sequel** Binary variable equal to 1 if film genre is tagged with sequel, meaning that one or more films were previously released in a series of films
13. **Genre_Thriller** Binary variable equal to 1 if film genre is tagged with thriller
14. **Production_Budget** Production budget in dollars
15. **Film_Runtime** Length of film in minutes
16. **Summer_May_Aug** Binary variable equal to 1 if film theatrical release date is in May, June, July or August
17. **March_Spring_Break** Binary variable equal to 1 if film theatrical release date is in March
18. **Valentines** Binary variable equal to 1 if film theatrical release date is between February 8 and 14
19. **October_Halloween** Binary variable equal to 1 if film theatrical release date is in October
20. **Winter_Holidays** Binary variable equal to 1 if film theatrical release date is between Dec 24 and Jan 2
21. **Rating_14A** Binary variable equal to 1 if the film is rated as 14A in the study geography, meaning that it's suitable for audiences 14 years or age and older
22. **Rating_18A** Binary variable equal to 1 if the film is rated as 18A in the study geography, meaning that it's suitable for audiences 18 years or age and older
23. **Rating_PG** Binary variable equal to 1 if the film is rated as PG in the study geography, meaning that it's suitable for all audiences but parental guidance is advised
24. **Award_Level_0** Binary variable equal to 1 if the film does not have any award nominations
25. **Award_Level_1** Binary variable equal to 1 if the film has 1-5 award nominations
26. **Award_Level_2** Binary variable equal to 1 if the film has 6-24 award nominations
27. **Big_Star_Count** Count of actors in the film who are in the top 100 films ranked by gross revenue for the study period

A data analysis workflow is built using the Alteryx Analytics platform and R programming language. A linear regression approach is utilized to predict the appeal of movie features for each dependent variable in Figure 1, resulting in a total of 11 models. A linear regression approach was chosen to present detailed results, as it offers a simplified output for interpretation of the significance and contribution of the predictor variables versus more advanced methods like neural networks. The initial regression equation is the same for each model and detailed in Equation 1.

$$Y = \alpha + \beta_1 \textit{Genre\_Adaptation} + \beta_2 \textit{Genre\_Adventure} + \beta_3 \textit{Genre\_Comedy} + \beta_4 \textit{Genre\_Drama} + \beta_5 \textit{Genre\_Family} + \beta_6 \textit{Genre\_Horror} + \beta_7 \textit{Genre\_Period} + \beta_7 \textit{Genre\_Romance} + \beta_9 \textit{Genre\_RomCom} + \beta_{10} \textit{Genre\_Sci\_Fi} + \beta_{11} \textit{Genre\_Sequel} + \beta_{12} \textit{Genre\_Thriller} + \beta_{13} \textit{Production\_Budget} + \beta_{14} \textit{Film\_Runtime} + \beta_{15} \textit{Summer\_May\_Aug} + \beta_{16} \textit{March\_Spring\_Break} + \beta_{17} \textit{Valentines} + \beta_{18} \textit{October\_Halloween} + \beta_{19} \textit{Winter\_Holidays} + \beta_{20} \textit{Rating\_14A} + \beta_{21} \textit{Rating\_18A} + \beta_{22} \textit{Rating\_PG} + \beta_{23} \textit{Award\_Level\_0} + \beta_{24} \textit{Award\_Level\_1} + \beta_{25} \textit{Award\_Level\_2} + \beta_{26} \textit{Big\_Star\_Count}$$

Equation 1: Initial regression equation where Y represents the % of the loyalty cardholder audience for a given film

Detailed results presented in the next section show the final regression equations resulting from excluding predictors that are not significant at the 90% confidence level, which provides a unique regression equation for each customer segment.

EMPIRICAL TESTING & RESULTS

Across the 11 models adjusted R-square values range from 0.39 to 0.84; results are visualized in Figure 1. The top performing model is Parents and the bottom performing model is Females 14-18. Regression equations for each of the 11 models are presented in this section with an analysis of the results offering further insight into the demographic cohort of moviegoers. The results provided in this section are extracted from Figure 2.

Results by Demographic Model

Parents/Guardians

The R Square for the parent segment is 0.84, indicating that the predictors used in this model explain most of the variance in the proportion of the audience that is made up of parents. Film ratings are more explanatory than genre variables for parents, the ratings 18A and 14A have negative coefficients, indicating that parents are not likely to take kids to films with these ratings. The genres Family, Sequel & Comedy yield positive coefficients, and indicate that parents are more likely to see film tagged with any of these genres versus films that do not have these genres. The regression equation is presented below in Equation 2.

$$\text{Prop\_Parents} = \alpha + \beta_1\text{Rating\_14A} + \beta_2\text{Rating\_18A} + \beta_3\text{Genre\_Comedy} + \beta_4\text{Genre\_Family} + \beta_5\text{Genre\_Sequel} + \beta_6\text{Genre\_RomCom}$$

Equation 2: Regression Equation for Proportion Parents Model

14-18 Age Cohort

The R Square for the female segment in the 14-18 age cohort group is 0.34 while the R Square for the male segment is 0.49, indicating that the predictors tested in this study explain much more of the variance for males in this age cohort versus females. For females, positive coefficients that indicate appeal are found for Horror, Romance, RomCom, March Spring Break and Award Level 0. For males, positive coefficients that indicate appeal are found for Horror, Action, Rating 18A, Award Level 0 and production budget. The regression equations are presented below in Equation 3.

$$\text{Prop\_F14\_18} = \alpha + \beta_1\text{Genre\_Action} + \beta_2\text{Genre\_Family} + \beta_3\text{Genre\_Period} + \beta_4\text{Genre\_Romance} + \beta_5\text{Genre\_Horror} + \beta_6\text{Genre\_RomCom} + \beta_7\text{Award\_Level\_0} + \beta_8\text{Summer\_May\_Aug} + \beta_9\text{March\_Spring\_Break} + \beta_{10}\text{Production.Budget}$$

$$\text{Prop\_M14\_18} = \alpha + \beta_1\ \text{Rating\_18A} + \beta_2\text{Genre\_Action} + \beta_3\text{Genre\_Adaptation} + \beta_4\text{Genre\_Drama} + \beta_5\text{Genre\_Family} + \beta_6\text{Genre\_Horror} + \beta_7\text{Award\_Level\_0} + \beta_8\text{Production.Budget}$$

Equation 3: Regression Equations for Proportion Males & Females Aged 14-18 Model

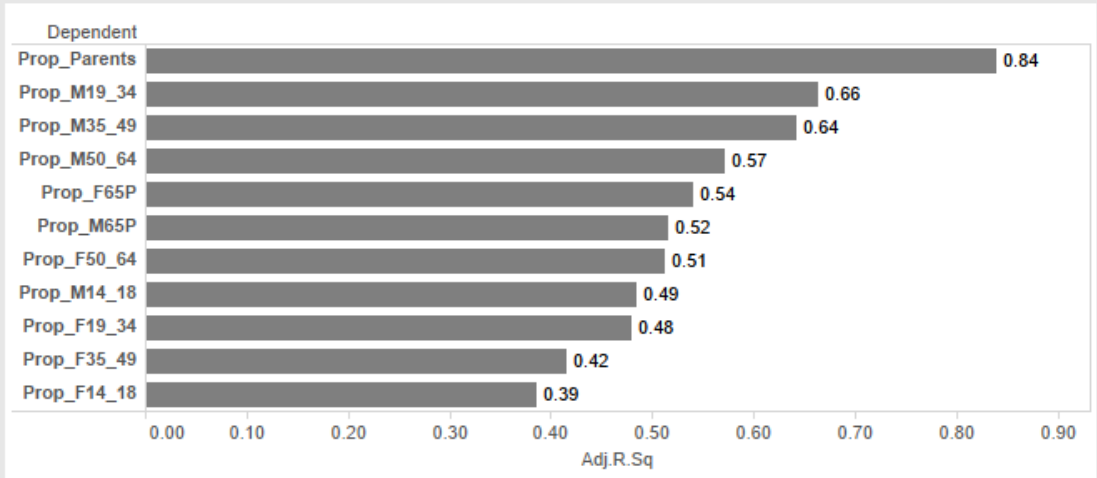
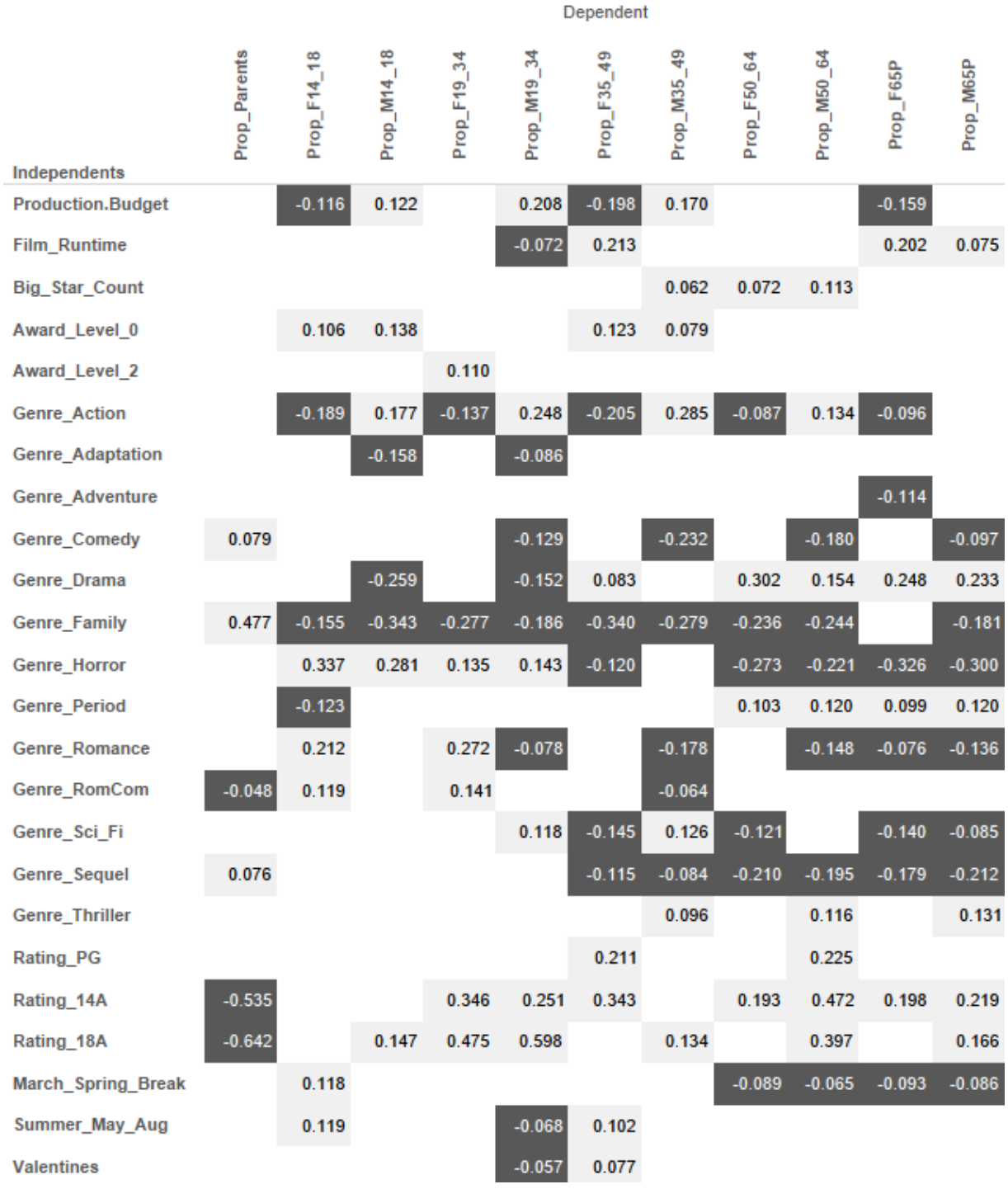


Figure 1: Adjusted R Square by Customer Segment Model



	Dependent										
Independents	Prop_Parents	Prop_F14_18	Prop_M14_18	Prop_F19_34	Prop_M19_34	Prop_F35_49	Prop_M35_49	Prop_F50_64	Prop_M50_64	Prop_F65P	Prop_M65P
Production.Budget		-0.116	0.122		0.208	-0.198	0.170			-0.159	
Film_Runtime					-0.072	0.213				0.202	0.075
Big_Star_Count							0.062	0.072	0.113		
Award_Level_0		0.106	0.138			0.123	0.079				
Award_Level_2				0.110							
Genre_Action		-0.189	0.177	-0.137	0.248	-0.205	0.285	-0.087	0.134	-0.096	
Genre_Adaptation			-0.158		-0.086						
Genre_Adventure										-0.114	
Genre_Comedy	0.079				-0.129		-0.232		-0.180		-0.097
Genre_Drama			-0.259		-0.152	0.083		0.302	0.154	0.248	0.233
Genre_Family	0.477	-0.155	-0.343	-0.277	-0.186	-0.340	-0.279	-0.236	-0.244		-0.181
Genre_Horror		0.337	0.281	0.135	0.143	-0.120		-0.273	-0.221	-0.326	-0.300
Genre_Period		-0.123						0.103	0.120	0.099	0.120
Genre_Romance		0.212		0.272	-0.078		-0.178		-0.148	-0.076	-0.136
Genre_RomCom	-0.048	0.119		0.141			-0.064				
Genre_Sci_Fi					0.118	-0.145	0.126	-0.121		-0.140	-0.085
Genre_Sequel	0.076					-0.115	-0.084	-0.210	-0.195	-0.179	-0.212
Genre_Thriller							0.096		0.116		0.131
Rating_PG						0.211			0.225		
Rating_14A	-0.535			0.346	0.251	0.343		0.193	0.472	0.198	0.219
Rating_18A	-0.642		0.147	0.475	0.598		0.134		0.397		0.166
March_Spring_Break		0.118						-0.089	-0.065	-0.093	-0.086
Summer_May_Aug		0.119			-0.068	0.102					
Valentines					-0.057	0.077					

Figure 2: Standardized coefficients significant at the 90% confidence level for independent variables across 11 customer segment models

19-34 Age Cohort

The R Square for the female segment in the 19-34 age cohort group is 0.48 while the R Square for the male segment is 0.66, indicating that, similarly to the 14-18 age cohort, the predictors tested in this study explain much more of the variance for males in this age cohort versus females. For females, positive coefficients are found for Rating 18A, Rating 14A, Romance, RomCom, Horror and Award Level 2. For males, positive coefficients are found for Rating 18A, Rating 14A, Action, Production Budget, Horror and Sci Fi. For males, positive coefficients are found for Rating 18A, Rating 14A, Action, Production Budget, Horror and Sci Fi. The regression equations are presented below in Equation 4.

$$\begin{aligned}\text{Prop\_F19\_34} = {} & \alpha + \beta_1\text{Rating\_14A} + \beta_2\text{Rating\_18A} + \\ & \beta_3\text{Genre\_Action} + \beta_4\text{Genre\_Family} + \\ & \beta_5\text{Genre\_Romance} + \beta_6\text{Genre\_Horror} + \\ & \beta_7\text{Genre\_RomCom} + \beta_8\text{Award\_Level\_2}\end{aligned}$$

$$\begin{aligned}\text{Prop\_M19\_34} = {} & \alpha + \beta_1\text{Film\_Runtime} + \\ & \beta_2\text{Rating\_14A} + \beta_3\text{Rating\_18A} + \\ & \beta_4\text{Genre\_Action} + \beta_5\text{Genre\_Adaptation} + \\ & \beta_6\text{Genre\_Comedy} + \beta_7\text{Genre\_Drama} + \\ & \beta_8\text{Genre\_Family} + \beta_9\text{Genre\_Romance} + \\ & \beta_{10}\text{Genre\_Sci\_Fi} + \beta_{11}\text{Genre\_Horror} + \\ & \beta_{12}\text{Summer\_May\_Aug} + \beta_{13}\text{Valentines} + \\ & \beta_{14}\text{Production.Budget}\end{aligned}$$

Equation 4: Regression Equations for Proportion Males and Females Aged 19-34 Model

35-49 Age Cohort

The R Square for the female segment in the 35-49 age cohort group is 0.42 while the R Square for the male segment is 0.64, indicating that, similarly to the younger age cohorts, the predictors tested in this study explain more of the variance for males in this age cohort versus females. For females, positive coefficients are found for Rating 14A, Film Runtime, Rating PG, Award Level 0, Summer, Drama and Valentines. For males, positive coefficients are found for Action, Production Budget, Rating 18A, Sci Fi, Thriller, Award Level and Big Star Count. The regression equations are presented below in Equation 5.

$$\begin{aligned}\text{Prop\_F35\_49} = {} & \alpha + \beta_1\text{Film\_Runtime} + \\ & \beta_2\text{Rating\_PG} + \beta_3\text{Rating\_14A} + \\ & \beta_4\text{Genre\_Action} + \beta_5\text{Genre\_Drama} + \\ & \beta_6\text{Genre\_Family} + \beta_7\text{Genre\_Sequel} + \\ & \beta_8\text{Genre\_Sci\_Fi} + \beta_9\text{Genre\_Horror} + \\ & \beta_{10}\text{Award\_Level\_0} + \beta_{11}\text{Summer\_May\_Aug} + \\ & \beta_{12}\text{Valentines} + \beta_{13}\text{Production.Budget}\end{aligned}$$

$$\begin{aligned}\text{Prop\_M35\_49} = {} & \alpha + \beta_1\text{Rating\_18A} + \\ & \beta_2\text{Genre\_Action} + \beta_3\text{Genre\_Comedy} + \\ & \beta_4\text{Genre\_Family} + \beta_5\text{Genre\_Romance} + \\ & \beta_6\text{Genre\_Sequel} + \beta_7\text{Genre\_Thriller} + \\ & \beta_8\text{Genre\_Sci\_Fi} + \beta_9\text{Genre\_RomCom} + \\ & \beta_{10}\text{Award\_Level\_0} + \beta_{11}\text{Big\_Star\_Count} + \\ & \beta_{12}\text{Production.Budget}\end{aligned}$$

Equation 5: Regression Equations for Proportion Males and Females Aged 35-49 Model

50-64 Age Cohort

The R Square for the female segment in the 50-64 age cohort group is 0.51 while the R Square for the male segment is 0.57. Similarly to the younger age cohorts, the predictors tested in this study explain more of the variance for males in this age cohort versus females, but the male-female difference in variance explained is much smaller than younger age cohorts. For females, positive coefficients are found for Drama, Rating 14A, Period and Big Star Count. For males, positive coefficients are found for Rating 14A, Rating 18A, Rating PG, Drama, Action, Period, Thriller and Big Star Count. The regression equations are presented below in Equation 6.

$$\begin{aligned}\text{Prop\_F50\_64} = {} & \alpha + \beta_1\text{Rating\_14A} + \\ & \beta_2\text{Genre\_Action} + \beta_3\text{Genre\_Drama} + \\ & \beta_4\text{Genre\_Family} + \beta_5\text{Genre\_Period} + \\ & \beta_6\text{Genre\_Sequel} + \beta_7\text{Genre\_Sci\_Fi} + \\ & \beta_8\text{Genre\_Horror} + \beta_9\text{Big\_Star\_Count} + \\ & \beta_{10}\text{March\_Spring\_Break}\end{aligned}$$

$$\begin{aligned}\text{Prop\_M50\_64} = {} & \alpha + \beta_1\text{Rating\_PG} + \beta_2\text{Rating\_14A} + \\ & \beta_3\text{Rating\_18A} + \beta_4\text{Genre\_Action} + \\ & \beta_5\text{Genre\_Comedy} + \beta_6\text{Genre\_Drama} + \\ & \beta_7\text{Genre\_Family} + \beta_8\text{Genre\_Period} + \\ & \beta_9\text{Genre\_Romance} + \beta_{10}\text{Genre\_Sequel} + \\ & \beta_{11}\text{Genre\_Thriller} + \beta_{12}\text{Genre\_Horror} + \\ & \beta_{13}\text{Big\_Star\_Count} + \beta_{14}\text{March\_Spring\_Break}\end{aligned}$$

Equation 6: Regression Equations for Proportion Males and Females Aged 50-64 Model

65+ Age Cohort

The R Square for the female segment in the 65+ age cohort group is 0.54 while the R Square for the male segment is 0.52, indicating that the predictors tested in this study explain only slightly more of the variance for females in this age cohort versus males. For females, positive coefficients are found for Drama, Film Runtime,

Rating 14A and Period. For males, positive coefficients are found for Drama, Rating 14A, Rating 18A, Thriller, Period and Film Runtime. The regression equations are presented below in Equation 7.

$$\begin{aligned}\text{Prop\_F65P} = {} & \alpha + \beta_1\text{Film\_Runtime} + \beta_2\text{Rating\_14A} + \\ & \beta_3\text{Genre\_Action} + \beta_4\text{Genre\_Adventure} + \\ & \beta_5\text{Genre\_Drama} + \beta_6\text{Genre\_Period} + \\ & \beta_7\text{Genre\_Romance} + \beta_8\text{Genre\_Sequel} + \\ & \beta_9\text{Genre\_Sci\_Fi} + \beta_{10}\text{Genre\_Horror} + \\ & \beta_{11}\text{March\_Spring\_Break} + \beta_{12}\text{Production.Budget}\end{aligned}$$

$$\begin{aligned}\text{Prop\_M65P} = {} & \alpha + \beta_1\text{Film\_Runtime} + \\ & \beta_2\text{Rating\_14A} + \beta_3\text{Rating\_18A} + \\ & \beta_4\text{Genre\_Comedy} + \beta_5\text{Genre\_Drama} + \\ & \beta_6\text{Genre\_Family} + \beta_7\text{Genre\_Period} + \\ & \beta_8\text{Genre\_Romance} + \beta_9\text{Genre\_Sequel} + \\ & \beta_{10}\text{Genre\_Thriller} + \beta_{11}\text{Genre\_Sci\_Fi} + \\ & \beta_{12}\text{Genre\_Horror} + \beta_{13}\text{March\_Spring\_Break}\end{aligned}$$

Equation 7: Regression Equations for Proportion Males and Females Aged 65+ Model

Model Results by Independent Variables

Figure 2 provides a comparative view of the independent predictor variables used by model. Blank cells in Figure 2 indicate that the independent variable was not significant at the 90% confidence level for the respective model. We summarize the results by drawing similarities and differences between the demographic cohorts in this section.

Production Budget has positive coefficients for males aged 14-49, indicating that males are more likely to see films with a bigger budget.

Film Runtime has positive coefficients for females 35 to 49 and both genders 65 plus. Males 19-34 have a negative coefficient on film runtime, indicating a preference for shorter films.

Big Star Count was only significant for three cohorts: both genders aged 50 to 64 and males 35 to 49.

The categorical **award nomination variables** indicate that films with no nominations appeal to teens and ages 35-49 while the opposite is true for females aged 19-34.

Action, where significant, shows positive coefficients for males and negative coefficients for females. With the exception of Females 65 plus, Romance shows the opposite of Action with the models for female segments having positive coefficients and males having negative, this is also comparable to the results from Redondo & Holbrook [9]. The genre **RomCom** appeals to younger females 14-34.

Drama, where significant, yields negative coefficients for younger demographics and positive coefficients for mature demographics. This is similar to the results from Redondo & Holbrook [9] where the fourth dimension indicated that a more mature audience was positively correlated with drama.

Across all but one model, **Family** was significant and offered strong explanatory power compared to most other independent variables. As expected, the parent cohort is likely to see Family movies, and all other age cohorts are unlikely to see movies tagged with Family. This finding is comparable with the results from Redondo & Holbrook [9] where the Family factor had the greatest model contribution and most distinct segment.

Horror has positive coefficients for younger cohorts and negative coefficients for older cohorts. **Period** appeals to a mature crowd with positive coefficients for demographics aged 50 plus. **Sci Fi** appeals to males 19-49 and Thriller appeals to males 35 and up.

Rating 18A is significant and positively correlated to all male age cohorts as well as females 19 to 34. **Rating 14A** is significant and has positive coefficients for all female age cohorts except 14 to 18 year olds.

The temporal variables **March Spring Break** and **Summer May Aug** are significant and have positive coefficients for high school aged females 14 to 18. **Summer May Aug** also has a positive coefficient for females aged 35 to 49. **Valentine's Day** has a positive coefficient for females aged 35 to 49.

CONCLUSION: DISCUSSION AND PRACTICAL IMPLICATIONS

This study explores methods for predicting audience composition and key customer segments for new release movies and proposes a new method using transactional customer loyalty data from a major movie exhibitor. The method proposed here uses a much larger behavioral dataset that offers several benefits versus the methods covered in the existing academic literature which utilize attitudinal survey data. Time to insight and cost are obvious advantages of this method, and it's possible to get results in seconds using this method versus any amount of time that would be involved with conducting consumer surveys. More granular insights are another advantage, and with a large enough transactional database, it's possible to produce more predictions for more films, such as how the key audience would vary between geographic regions, day of release or time of day. The big advantage for organizations having a

customer loyalty program is that this method allows for programmatic use of these predictions in internal systems. Many of the decisions that can be informed based on this data require some type of blending of datasets. Operational decisions around film and screen booking require the key customer segment predictions be compared with location-specific customer attributes. To inform affiliate advertising with complementary products, key customer segment predictions must be compared with product-specific customer attributes. For large organizations, there may be hundreds or thousands of locations or products, which would be too much data for managers to sift through manually. Therefore, programmatic methods would enable the most efficient use of customer segment predictions. For example, the key customer segment predictions could inform recommendation engines for managers that suggest how to allocate movies to screens and locations or which food products to pair with film-based advertising.

At the time of this paper, there were no academic studies found that produce demographic audience composition forecasts and key segments by film as described in the following section, named 3: Data & Method. Moreover, the existing studies that examine moviegoer demographics quantitatively primarily utilize customer survey data, such as that collected from a targeted research study and national census programs. This study demonstrates the application of behavioral customer loyalty data for forecasting demographics for new movies and offers an output format readily consumable by business managers in the movie industry. The approach of using loyalty data to predict audience composition for new products in the entertainment industry differs from a survey's data-based approach in its potential for scalability, time to insight and embedded analytics. Surveys can be costly to run, and without a large sample and response rate it can be difficult to get down to the granular segments of customers. Essential to practical applications, embedded analytics refers to decision support systems embedded in business processes and routinely utilized in the decision making process. To realize the value of predictive audience composition methods in marketing, advertising and operations for new entertainment products, the predictions must be served up in some form of automated fashion. Take for example the application of using audience composition predictions to inform affiliate advertising with complementary products. In the most simplistic form the automated process could consist of a dashboard displaying the predicted audience for new release products, which gets updated daily as the predictive model runs for new products. Since the products are added to a central database, this process only becomes valuable and embedded when a manager utilizes the dashboard on a regular basis to recommend complementary products to feature with the new release movie. With the same application of affiliate advertising for complementary products, an example of a more advanced application of embedded analytics would be a scheduled algorithm that sources the key audience for the affiliate products from a related database, compares them to the predictions for new movies and selects creative work and copy for digital menu boards that display movie-themed food bundles at the concession stand or online when customers have purchased their tickets. The latter example is one type of intelligent system offering operational efficiencies and an opportunity for targeted communications at a granular level. With a large enough transactional database, it's possible to produce more predictions for more products and get down to granular audience segmentation, such as by geographic region, locations, dates or time of day. For large organizations such a major movie exhibitors, there may be hundreds or thousands of locations or products. However, this would be too much data for managers to sift through manually, so programmatic methods enable the most efficient use of audience composition predictions.

The empirical results presented demonstrate promising evidence that customer loyalty data can be used to predict audience demographics for new release movies. The independent variables used are good predictors for most age cohorts with all models having an explanatory power over 39%, which is acceptable in the social sciences for gaining directional insights into the variables that impact the behavior of humans. These results are also directionally similar to the overall explanatory power of the model built by Redondo & Holbrook [9] with 45% of variance explained.

The large range in explanatory power across the 11 demographic models reveals an opportunity to source and test additional variables to generate more accurate predictions for certain demographic cohorts, particularly for younger female cohorts. Existing literature can provide direction on additional explanatory variables to research. For example, McKenzie and Walls' [6] results show that piracy has an adverse effect on film performance that is amplified when the local theatrical release date is later than the American theatrical release date. The days between release dates could be used as an independent variable in our model to provide directional insight to variation in how piracy impacts segments of moviegoers. The study by Westland [15] shows that social media data points can be used to predict box office revenue. Similarly, the volume of posts could be represented as an independent variable in our model to assess the impact of social media on various demographic segments.

Variables we use in this study that were not significant for many cohorts should be researched further using different independent variables. For example, we use the variable 'Big Star Count' to represent the appeal of actors in a film based on the number of actors in the film who are in the top 100 films ranked by gross revenue for the study period. We have not tested other ways of representing actors in the model and suggest a subsequent academic study to analyze different methods for representing and analyzing actor appeal to demographic cohorts.

We recommend further exploratory analysis to better understand the purchasing habits of loyalty card holders when they go to the movies. In our study we look specifically at the behavior of the loyalty card holder and do not consider the other people they are going to the movies with, this could be an influential factor on choice of movie. For example, the choice of movie for a couple on a date night could be different than the choice for a group of friends. An exciting area for future academic research is the adaptation of our approach using customer loyalty data to other types of audience segments. For example, a multivariate data reduction method could be used to define clusters of moviegoers based on who they are going to the movies with, or what the occasion is. Psychographic, lifestyle and behavioral attributes can all be incorporated. The percent of audience for resulting clusters could be used as a dependent variable in the model instead of the demographic segments our study uses. Moreover, this method can be applied to other new product releases in the entertainment industry, such as books, music or performing arts.

Lastly, in addition to testing and refining the independent variables in our models, we recommend further analysis to understand and validate how this model would perform in a real world setting. In our study we utilize a single dataset to build our regression models but do not test the models on a different dataset. Doing such an exercise could help with practical application and adoption of this method in a business setting. The awards nomination variables we use are ex-post and therefore will be problematic if the intended application is forecasting. We suggest dropping these variables from the model in this case.

The development of algorithms and embedded analytics is an exciting area of research with great opportunities for practical applications in organizations wanting to develop a competitive advantage using customer data and operations research methods.

REFERENCES

[1] Basuroy, S., Chatterjee, S. & Ravid, A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power and Budgets. *The Journal of Marketing*, 67(4), 103-117.

[2] De Vany, A. & Walls, D. (2002). Does Hollywood Make Too Many R-Rated Movies? Risk, Stochastic Dominance, and the Illusion of Expectation. *The Journal of Business*, 75(3), 425-451.

[3] Einav, L. (2007). Seasonality in the U.S. Motion Picture Industry. *The Rand Journal of Economics*, 38(1), 127-145.

[4] Eliashberg, J. & Shugan, S. (1997). Film Critics: Influencers or Predictors? *The Journal of Marketing*, 61(2), 68-78.

[5] Litman, B. (1983). Predicting Success of Theatrical Movies: An Empirical Study. *Journal of Popular Culture*, 16 (4), 159-175.

[6] McKenzie, J. & Walls, D. (2016). File Sharing and Film Revenues: Estimates of Sales Displacement at the Box Office. *Journal of Economic Analysis & Policy*, 16(1), 25-57.

[7] Medved, M. (1992). Hollywood vs. America: Popular Culture and the War on Traditional Values. HarperCollins. ISBN 0-06-016882-X.

[8] Ravid, S. (1999). Information, Blockbusters, and Stars: A Study of the Film Industry. *Journal of Business* 72(4), 463–492.

[9] Redondo, I. & Holbrook, M. (2010). Modeling the appeal of movie features to demographic segments of theatrical demand. *Journal of Cultural Economics*, 34(4), 299-315

[10] Sharda, R. & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications*, 30(2), 243–254

[11] Sawhney, M. & Eliashberg, J. (1996). A Parsimonious Model for Forecasting Gross Box Office Revenues of Motion Pictures. *Marketing Science*, 15(2), 113-131.

[12] Smith, S. & Smith, V. (1986). Successful Movies: A Preliminary Empirical Analysis. *Applied Economics*, *18*, 501-507.

[13] Somlo, B., Rajaram, K. & Ahmadi, R. (2011), Distribution Planning to Optimize Profits in the Motion Picture Industry. *Production and Operations Management*, 20(4), 618-636.

[14] Wallace, W. & Holbrook, M. (1993). The Role of Actors and Actresses in the Success of Films: How Much is a Movie Star Worth? *Journal of Cultural Economics*, 17(1), 1-27.

[15] Westland, J. (2012). The adoption of social networking technologies in cinema releases. *Information Technology Management* 13(3), 167–181.

[16] Zufryden, F. (1996). Linking Advertising to Box Office Performance of New Film Releases: A Marketing Planning Model. *Journal of Advertising Research*, 36(4), 29–42.

AUTHOR BIOGRAPHIES

Katherine Goff Inglis has been extracting business value from movie data since 2008 at Cineplex Entertainment where her current role is Director, Data Science and Analytics. She received her PhD (2017) in industrial engineering from Ryerson University in Toronto, Canada. She is specialized in operations research and has wide knowledge in enterprise analytics, data management, and loyalty reward programs. Katherine is also a Professor of Marketing Analytics at Centennial College in Toronto, Canada.

Saeed Zolfaghari is a professor of industrial engineering at Ryerson University in Toronto, Canada. He received his PhD (1997) from the University of Ottawa, Canada. His primary research interest includes productivity improvement of manufacturing and business operations through proper use of resources and reduction of waste. He is specialized in operations research and has a broad knowledge in the area mathematical modelling and metaheuristics. Other areas of his research include loyalty reward programs, forecasting, logistics, cellular manufacturing systems, and scheduling. He is a senior member of IIE and IEEE, and a member of CORS and PEO.