ACCT414: ACCOUNTING ANALYTICS CAPSTONE (SMU-X)

CATHAY CINEPLEXES

SMU SINGAPORE MANAGEMENT UNIVERSITY

Group 3

Angelo

Callista

Donald

Evonne

Pou I

Ruo Rong

Problem Statement

How can **Cathay Cineplexes** leverage **relevant data factors** in a predictive model to **estimate box office performance** to aid operations and decision-making.

Goals

Uncover the most **influential factors** (predictors) that **drive box office performance**.

Develop a **predictive model** to estimate box office for upcoming films, factoring in **historical data** and other **relevant characteristics.**

historical data + relevant external data → predict total '**Admits**'

Methodology

- Data Cleaning
- Variable Selection
- Categorical Weight Assignments
- Negative Binomial Regression
- Out-of-sample test
- Evaluate Prediction

Data Cleaning | Variable Selection

We begin with joining the different tables into one '**Transaction**' table/sheet based on the given primary keys.

Stratification: to group each observations by their Average Admits or Ratings (public data)

[redacted]

Weight Assignments: weights (numeric) are assigned to each observations - higher Average Admits / Ratings → bigger weights

[redacted]

Factoring in branch effect: standardized data across branched, and to take into account difference performance and preference within different branches.

Models & Accuracy

Negative Binomial is a better model to predict box office due to 'Total Admits' being **discrete** data, and **overdispersed** (film admits are very spread out). It models the log (Admits) while adjusting for the fact that some films may perform way better/worse.

R Studio

Trim 25% of data

We decided to trim our training data in hope to remove unnecessary outlier and get better prediction accuracy

Model 1 → the one with interactions

Using Stepwise selection to come up with the optimal set of variables along with its interactions.

GenreWeight: LangWeight + DirectorWeight: DistributorWeight + ...

Model 2 → the one with more factor variables

Aim to retain some multi-level factors variables which are hard to do interactions with. Weight variables are still included.

Out-of-sample test:

[redacted]

Key Findings

Overpredictions tend to occur for films with lower actual admits.

Underpredictions become more significant as the number of actual admits increases.

This **limited sensitivity to extreme box office successes**, means bigger gaps between actual and predicted values as the actual number gets lower/higher.

Models tend to do very well on **1000 - 6000 admits** range.

Things to note:

- limited access to ratings (public data)
- unpredictable market trends
- the models are trained on historical data and international rating data, need to account for local market preferences

Predictor Variables

Variables are selected based on research papers & preliminary OLS regressions.

Film Opening Date	Film Duration	Genre	Censorship	Cinema Branch
Opening Year | Month | Quarter	Film Duration (Hours)	Genre 1 |2|3 | Genre Weight	Film Censor | Censor Weight	Branch No. (Factor)
Actor	Distributor	Language	Director	
Tier 1 | 2 | 3 | avg Actor Rating	Distributor Weight	Language | Language Weight	Director Names | Weight	

Weights assigned on average admits

Weights assigned on average ratings

Recommendations

- **Standardizing key variables** for consistency and improving interpretability
- While historical data provides a strong foundation, **incorporating customer preferences**, current local market trends, and environmental factors could further refine predictions.
- **Addressing extreme anomalies** to improve robustness and **expanding data sources**, could provide a competitive edge in forecasting accuracy, given the dynamic nature of box office performance.

Application Example

Link to Power App

[redacted]

Application of Model

Gather upcoming film details → Input results into platform → To determine 'total tickets predicted' in the app → Make relevant business decisions