Region Based Box Office Revenue Prediction for Telugu Movies

Problem Statement:

The prediction of box office revenue on the basis of information that is known very early on in the production of the movie can be extremely beneficial to producers, directors, and actors as it can assist in determining the budget of the movie and whether it is a lucrative prospective or not. In the Telugu film industry in particular, there is an assumption that star power is the only factor that contributes to the success of the film. I have also noticed that people sometimes watch movies because they come from the same region as the actor and am interested if this is a prevailing sentiment. Knowing whether this is true can also assist in marketing the movie to the right audience. A majority of my family and cousins also decide whether to watch a movie or not based on the music composer. Hence, I wanted to assess whether these assumptions are true and if music also plays a large role in box office success.

Goal:

My goal was to create a model that can predict the total Box Office Revenue of Telugu Movies and can also reveal the importance of factors such as star power play in predicting the same. In my model, I also wanted to test whether certain actors are truly more popular in certain regions. Hence I decided to create a model that not only predicts the total Box Office data, but also predicts the box office performance in each region based on factors such as the cast, producer, directors, and composers of the movie.

Necessary Background:

There are 8 box office regions that together make up Andhra Pradesh and Telangana. They are Nizam , Ceded, Vizag(or Uttarandhra), East Godavari, West Godavari, Krishna, Guntur, and Nellore. The Nizam region makes up most of Telangana, while the rest belong to Andhra Pradesh. We will be using box office share data and not box office gross data.

Data Sources and Collection:

1. **IMDb**: I built a scraper using the BeautifulSoup library that could extract the name, directors, writers, producers, music composers, plot synopsis, release date, and the top 5 cast members of a movie. I used it to scrape this data for movies from 2010 to 2024 which resulted in a list of around 2500 movies.

   I ran the scraping code separately for each year of movies instead of providing all the urls to every movie all at once as I wanted to be able to address any errors quickly. The scraping programs also took quite a while to run and I did not want to waste the time my computer had spent on producing the rest of the data due to a small error somewhere. After the program created one excel sheet for each year, I saved it all in one file.

2. Box Office Andhra: I manually copied the regional data provided in this website into excel and this data formed the basis of my model. Although it only had regional box office data for around 140 movies originally made in Telugu, it was the only website that

provided a clear list of such data. Scraping was not possible as there was no pattern to the way the website structured its pages.

3. Wikipedia and Telugu Movie News Websites: After scraping the IMDb data, I tried to find the total box office revenue for as many of the scraped movies as possible. I used wikipedia and a variety of Telugu Movie News Websites to manually find information on the same.

Methodology:

1. Cleaning and Preparing Data: I scraped all information except box office information for all movies between 2010 and 2024, then deleted any movies that had empty values. Next, I manually collected regional box office share data for as many of these movies as possible and matched this data to the scraped data.
2. One-Hot Encoding: After having scraped and collected all the necessary information, I had lists of the top 5 movies, lists of directors, and lists of producers. To tackle this, I implemented one-hot-encoding on all of these variables, turning them each into features of their own. This resulted in me having nearly 560 features as opposed to the 140 data samples I had. Hence, I picked the 40 features that appeared the most number of times to use. These included producers, month of release, directors, actors, and composers.
3. Lasso Regression: I then implemented linear regression using the lasso technique. My data was quite sparse and feature selection was incredibly important, so I chose this technique to optimise my results. I implemented lasso regression using Python's sklearn library. I split my data 80:20 into 20 into training and test and ran a lasso regression separately for each region, using different alpha values for each.

   I decided on these alpha values after trying a variety of them and assessing which resulted in a balance between penalising unnecessary features and not reducing the number of predicting features so much that the predictions lacked diversity across movies.

4. Result Visualization: Once I had my coefficients and my predicted data, I downloaded it into excel sheets to better visualize it. I then arranged the coefficients in descending order for each region to see what features contributed the most to the box office revenue

Challenges and Shortcomings:

My plan initially involved finding the regional data for around 1000 of the 2500 movies I scraped from IMDb, however I was unable to do so. Each movie required me to personally scour through websites hoping to find the box office data and inputting it in an excel spreadsheet. In most cases it was impossible to find the total box office data, let alone regional divisions in the revenue. Hence, I decided to stick to the regional data I had sourced from the Box Office Andhra website despite it being a very small dataset to run a machine learning model on.

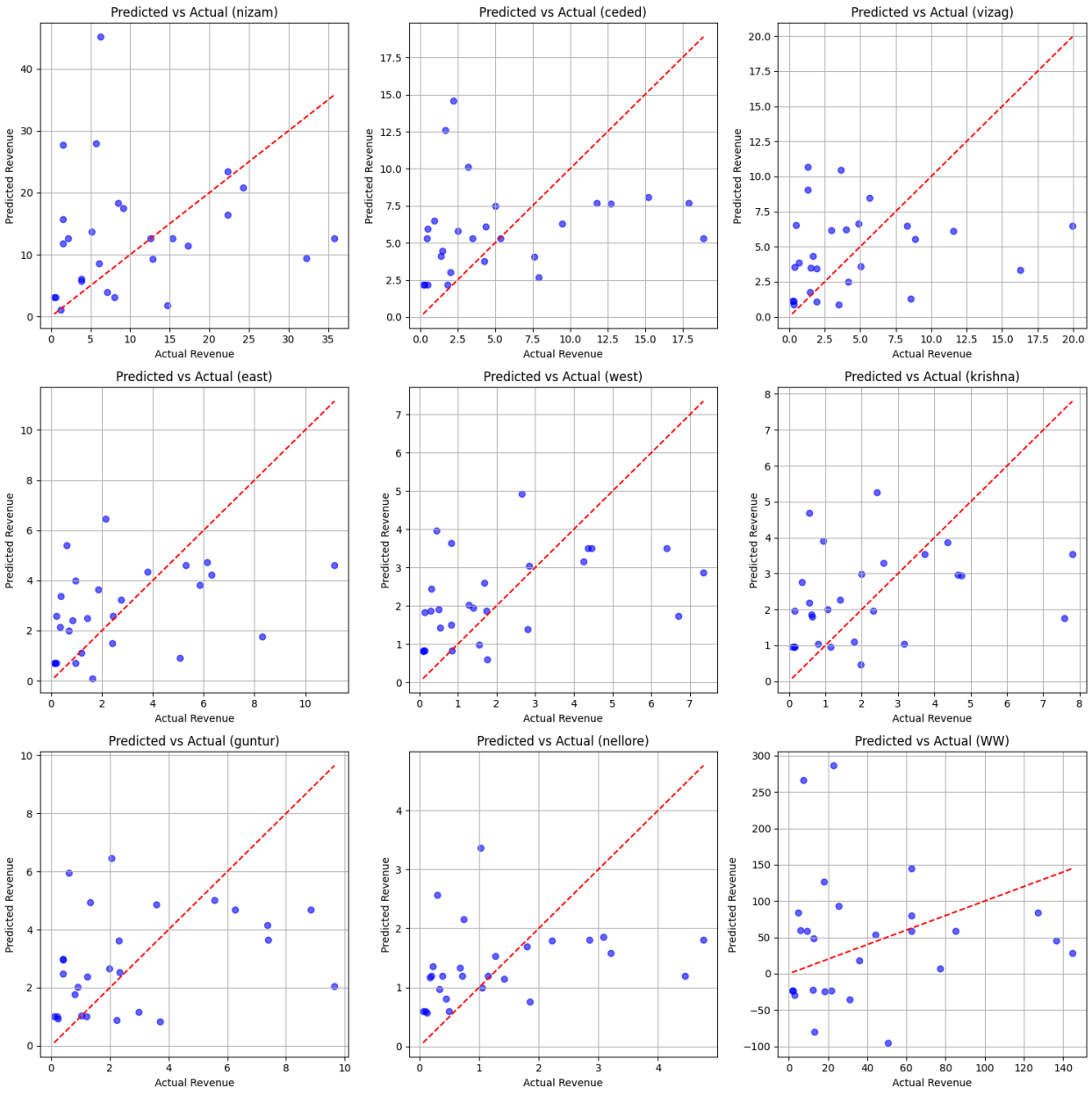
I had also intended to include budgetary data so that it would be more apparent which movies were actually profiting as opposed to earning a lot but spending a lot more. However, I faced a

data availability issue. A majority of movies had no information anywhere regarding their budget. Hence, I discarded my idea to include this as a feature.

Furthermore, a majority of the data that I have been able to find is data related to very high budget, successful movies that star very popular actors and actresses. Hence, there is an issue with predicting mid-budget and low-budget movies. The data available inherently reflects the perception that most Telugu audiences care the most about the actors in the movie rather than the quality of the movie itself. Hence, I knew it was likely that it would affirm the positive relationship between big stars and box office revenue.

Results:

The small size of my samples and my vast number of features did prove to be a problem, as the resulting box office revenue predictions were not very accurate. However, the inferences from examining the coefficients of each feature are intriguing. The coefficients are available in the coefficients_analysis.xlsx spreadsheet.



1. In every region, Prabhas, and N.T. Rama Rao Jr. remain at the top, followed closely by Ram Charan, Mahesh Babu, Allu Arjun, and M.M. Keeravani. While it does not rule out the possibility that an actors' regional background could determine the success of a movie, it does provide some evidence against it. The same actors, the ones who are currently the most successful and popular in the industry, contribute to success in different regional markets regardless of their background. However it is interesting to note that Chiranjeevi, who hails from West Godavari, has the highest contribution towards predicting the box office in West Godavari.
2. It is also worthwhile to note that the worldwide collections show the greatest variation from the trend of big stars being the only ones at the top. Anil Ravipudi(a director) and Rashmika Mandanna(an actress) are near the top, which indicates that worldwide audiences may be more inclined than local audiences to watch movies that aren't boosted by their star power.
3. From the data, it does seem that the actors are the determining factor when it comes to predicting box office data, however as mentioned earlier, the data is inherently biased.
4. With regards to music, the high positive correlation that composers M.M. Keeravani and S. Thaman show in all regions underlines the importance of solid music to people who are buying tickets.

Further Work:

1. Building a classification model that determines whether a film will be a blockbuster, super hit, hit, average, flop, or disaster. These terms are directly related to box office success and finding data related to this is far easier.
2. Finding more diverse data that has information about lower budget movies as well would lead to more accurate predictions.
3. Introducing engineered features such as star power, which can be created by averaging the average box office revenue for all the movies the main cast of a particular movie has acted in. Creating a feature that represents pairings between directors and actors may also lead to more interesting results.

Github Repository Link:

https://github.com/Lahari909/IML_Final_Project_Lahari