STAT 426 Final Project — Proposal

Isabelle Tooley
November 20, 2020

1. What is the main question that you are trying to answer? Be as specific and detailed as possible.

The main research question for my project is to see if it is possible to accurately predict the primary genre of a full-length movie.

2. Where are you getting your data? What does your data look like?

The data are from a combination of three datasets. Two of them are from the IMDb website: a "basics" dataset with general information about the film content including ID, title, runtime, and genre; and a "ratings" dataset with identical IDs, average rating, and number of votes. The third dataset is from Kaggle and has IMDb IDs and the plot in text format, which was collected from the IMDb site and Wikipedia.

What is the target?

The target variable is the main genre of the film. After exploring the data, I have decided to limit the possible classes from 20 to 7: **action, comedy, drama, crime, horror, adventure,** and **other** (which is a lumped group of all other less common genres). This is to reduce the possible affects of class imbalance in the data (e.g. 2800 action movies vs 7 history movies).

What are the features?

The features that I have narrowed it down to are:
- startYear
- runtimeMinutes
- averageRating (the weighted average of user ratings of the movie)
- numVotes (the number of user votes for the movie)
- plot_synopsis (plain text)
- **genre**

	tconst	primaryTitle	startYear	runtimeMinutes	averageRating	numVotes	plot_synopsis	genre
0	tt0002130	Dante's Inferno	1911	71.0	7.1	2293	The exhumation of Lizzie Siddal's desiccated b...	Adventure
1	tt0003419	The Student of Prague	1913	85.0	6.5	1799	Being praised as the finest fencer in his Univ...	Drama
2	tt0003489	The Last Days of Pompeii	1913	88.0	6.2	475	In Pompeii 79AD, Glaucus and Jone are in love ...	Adventure
3	tt0004022	Julius Caesar	1914	112.0	6.2	42	The play opens with the commoners of Rome cele...	Drama
4	tt0004099	The New Wizard of Oz	1914	59.0	5.2	418	King Krewl (Raymond Russell) is a cruel dictat...	Adventure
5	tt0004635	The Squaw Man	1914	74.0	5.7	883	James Wynnegate (Dustin Farnum) and his cousin...	Action
6	tt0004972	The Birth of a Nation	1915	195.0	6.3	22474	=== Part 1: Civil War of United States ===\nTh...	Drama
7	tt0005059	The Captive	1915	50.0	6.4	77	The Captive chronicles the life of a young wom...	Drama
8	tt0006206	Les vampires	1915	421.0	7.3	4213	=== Episode 1 – "The Severed Head" ===\nPhilip...	Action
9	tt0006780	Hell's Hinges	1916	64.0	6.8	845	Hell's Hinges tells the story of a weak-willed...	Other

the above table is the result of multiple merges, feature selection, and some cleaning

Are you planning on creating / engineering new features?

The only features I plan on creating are TF-IDF scores for the "plot_synopsis" text content. I haven't looked into it yet, but I plan on limiting it to terms that are not extremely rare within the entire context of the data, but also not extremely common (not including stop words). Before doing a TF-IDF vector transformation, I'll also need to go through and do some cleaning of the text.

How many observations do you / will you have?

After merging all of the tables and handling missing values, I am left with a **total of 12,259 observations**, which seems like a decent amount of data. It is enough to be able to train the model well, but hopefully not too much to the point of extensive computation time.

How does the data that you are collecting help you answer your question of interest?

First of all, I do think that I'm going to remove the "tconst" and "primaryTitle" variables. These are unique to each observation and will provide no value in modeling. I might also remove the "startYear" feature because I would treat that as a factor instead of numeric variable, and that would add quite a few features (considering that each level will have its own dummy variable). And I'm not confident in the value that the release year of the movie would add to the model.

The rest of the features ("runtimeMinutes", "averageRating", "numVotes", and "plot_synopsis") I see being very helpful to making genre predictions. When it comes to ratings, it could be that certain genres are more popular among viewers than others. Runtime could also differ between different types of movies. As an example of this, although not necessarily relevant to this data, animated family films are typically shorter than more action-based movies. The plot summary or synopsis is probably going to help the most in predicting the genre of the movie as the two are directly related. Genre is the structural form or style of a story, where movies within a genre always include similar characteristics or conventions of storytelling. Originally I want to use scripts, the synopsis is a condensed version of the movie content and main story. And because genre is determined by the story told, plot should (in theory) be a good predictor.

one concern I have about this project is that the genres are similar to each other. There might not be strong enough differences within the data to distinguish between action and adventure movies, or horror and crime movies. I don't think this makes the project a waste of time or fundamentally flawed… but it will definitely be interesting to see how well the data that I have lends itself to genre classification when the boundaries between classes aren't that extreme.