STOR 565 Project Proposal
Manhui Zhu; Chengze Xie

General Description: With the rapid development of streaming media, people spend more and more time on them for entertainment. Researching user preference and improving the user retention time are important to various streaming platforms. In this project, we have 2 tasks: 1) build several models to predict the popularity (rate) of movies 2) make the recommendation according to the research of users.

Datasets:

Data sources: Kaggle

Name	General Description
The Movies Dataset	It contains 6 files that include all 45,000 movies listed in the Full MovieLens Dataset. Data points include cast, crew, plot, keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has a file that includes 100,000 ratings from 700 users on 9,000 movies, which is the subset from Full MovieLens Dataset.
TMDB 5000 Movie Dataset	This dataset has two files, the first file "movies" has 20 variables on 4803 movies. The variables are *budget, genres, keywords, etc.* The second file "credits" has 4 variables on 4803 movies. The variables are *movie_id, title, cast,* and *crew*.

Technique:

In the first tasks, we will use both linear and nonlinear regression models to predict how different features related to the movies influence the rating of movies. First, we will use the IMDB formula to create the weighted rating variable as the response variable. For linear regression models, we will do Lasso & Ridge, and PCR & PLS to control the model complexity. For nonlinear regression models, we will use the spline model. For both linear or non-linear model, we will employ cross-validation techniques to choose the best tuning parameter.

In addition, we may try some recommendation machine learning techniques, such as Decision Tree Classifier, “e-learning recommendation system”, “e-learning recommender”, adaptive e-learning”, “recommendation system”, and “e-learning”, which are widely used in recommendation system field. The relative papers of corresponding method are here.

https://www.researchgate.net/profile/Divya-Bharathi-P/publication/319284117_Student_placement_analyzer_A_recommendation_system_using_machine_learning/links/60503776458515e8344a6d09/Student-placement-analyzer-A-recommendation-system-using-machine-learning.pdf

https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed

https://researchoutput.csu.edu.au/ws/portalfiles/portal/56945122/37059927_Published_article.pdf

Potential Challenge & Attacking Plan:

There are many challenges for us. For instance, among all recommendation we need to pick up some and learning how to implement each one them. We definitely need to read some papers to understand the ideas of algorithms and we also need to learning and find a fair way to test the method we implement. We prepare to spend one week to ten days to understand the general ideas and most useful and applicable theorems that we want to use, then finish coding part within one week, and in the rest of time, we will test and record all data we get (including hyperparamters and accuracy of different methods).