Machine Learning Project Proposal
“Movie Revenue Predictor”

Supakjeera Thanapaisal
Francisco Perales
Christian Bargraser

Outline

We will be using regression to predict movie revenue. We are focused on creating an optimal process for selecting and creating the best model for our regression task. People typically simply call ***train_test_split()*** and then test their models over ***n*** epochs; however, we want to go beyond this simple approach and use more complex techniques such as dimensionality reduction, grid search, and k-fold cross validation. We have chosen to take on a regression problem so that we can gain more experience with evaluating regression models. Evaluating performance of models for classification problems is rather straightforward since metrics such as precision, recall, and f1-score can be compared. However, it takes some more thought to compare RSME values, as it is not immediately obvious what a good RMSE value for the current task is.

The approach we will take is the following:

- Clean Data
- Feature Extraction
- Analyze and Encode Data
- Grid Search
- K-fold Cross Validation
- Model Selection

Dataset Description

Total Movies	7400 movies
Train	70%
Test	30%
All of the Columns:	id Belongs_to_collection Budget Genres Homepage Imdb_id

	Original_language Original_title Overview Popularity Poster_path Production_companies Production_countries Release_date Runtime Spoken_languages Status Tagline Title Revenue
Label:	Revenue
Selected Features (8):	id Budget (matters since some are cheap movies) Genres (matters, hero fiction $$$) Original_language (yes) Production_companies (matters = marvel) Production_countries (USA movies = famous?) Release_date (during halloween?) Runtime (duration of movie in minutes) Spoken_languages (yes)

Clean Data

We will start off by determining if we should drop rows with null values. If a small percentage of rows are lost, then dropping rows with null values is acceptable. If a large percentage of rows are lost, then we will consider replacing null values with the mean or median, for numerical and categorical features respectively. We will also drop columns that are irrelevant towards the regression task.

Feature Extraction

Once we have cleaned our data, we will attempt to perform feature extraction. We will see if it is possible to combine multiple features into new features that provide additional insight into the data.

Analyze/Encode Data

Once we have extracted our features, we will look to see if there are linear relationships between our numerical features. We will do this by looking at the Pearson correlation coefficient. If two features are highly correlated, we will drop one of the features. This is because if two features are highly correlated it means that one of the features is not contributing much to the model. Dropping redundant features will serve as a form of dimensionality reduction and reduce the risk of overfitting the data.

Encode Data

We will then encode the features. We will one-hot encode categorical features and use a standard scaler on numerical features. We could normalize the numerical features, but we have chosen to use a standard scaler since it is less sensitive to outliers.

Grid Search

We will train several different models. We are interested in comparing the performance of linear regression, SVM, random forest, naive bayes, knn, and potentially a neural network if we have time. However, simply training models using the default parameters for each model and then evaluating their performance is an arbitrary comparison of model performance. Since model performance can be greatly affected by hyperparameter tuning, we want to compare the models after finding the best hyperparameter settings for each model before comparing model performance.

K-fold cross validation

In simple settings, models are evaluated over some large ***n*** epochs. Although effective when a large enough n is used, we want to use the more efficient cross validation. We are unsure what number of folds we want to use. We could use leave-one-out cross validation, but that would likely take longer than k-fold cross validation.

Model Selection

We will select the model that has the lowest RSME.