thefinitemonkey / **udacity_mlcapstone**

Branch: **master** ▾ **udacity_mlcapstone** / **proposal.md** **Find file** **Copy path**

thefinitemonkey code: update proposal and add basic stats exploration to notebook 0ed8ad9 a minute ago

1 contributor

```
45 lines (29 sloc)   5.29 KB
```

Machine Learning Engineer Nanodegree

Capstone Proposal

Doug Brown
February 23, 2019

Proposal

(approx. 2-3 pages)

Domain Background

The proposed project focuses on the entertainment industry, specifically dealing with movies. The movie industry generated an estimated $41.7 billion in revenue for 2018. Investments in films is also high, with the budgets for blockbuster summer releases regularly surpassing $200-300 million (see https://www.the-numbers.com/movie/budgets/all).

Understanding how successful a film is likely to be based on factors such as cast, director, screenwriters and so on can help a studio make better investments in their production costs. Knowing where to draw lines to maximize profits can make for a healthier industry, and players such as Netflix have been employing their own viewership data to create projects such as House Of Cards. And as a consumer, having an idea of whether a movie will be worth spending time and money viewing isn't a bad thing either.

Problem Statement

Given a sample and test datasets describing key attributes of movies, create a system for predicting worldiwde revenue for a given movie release. The test dataset contains nearly 4400 movies against which predictions will be measured, and the training set includes 3000 movies.

Datasets and Inputs

The dataset for this project comes from The Movie Database (https://www.themoviedb.org/), a leading source of movie data. This data includes all the quantifiable factors about each movie, such as cast, crew, plot keywords, budget, release dates, languages, production companies, genres, etc. All data points contribute to the desirability of a movie, and therefore its earning potential. Some values may have individual impact, while others may be combinatorial.

The dataset is broken into two files already. One file is the training file with 3000 movies, and the other is the test file with 4398 movies. The test file does not contain the revenue data point. The model cannot be over-fit to the test data as a result.

The dataset and problem are both part of the Kaggle competition available at https://www.kaggle.com/c/tmdb-box-office-prediction/

Solution Statement

This problem will be solved using a Supervised Learning approach. This will allow for use of the historical, known data points and outcomes to create a model with which to predict unknown outcomes for new instances of similarly quantified data. The predictions made for worldwide revenue of each movie in the test dataset will be submitted and judged for accuracy against the known outcomes that are not included in the test data.

Benchmark Model

A Regression model is one that could potentially fit this problem space. Regression can fit this type of data very well, and all the data in this set can be made neatly numerical. Each data point is either already represented as a number, or can be transformed to a numeric representation. The decision tree involved here will be done on a higher number of dimensions due to the number of data fields, fitting worldwide revenue as the model output.

Evaluation Metrics

This project will be evaluated on Root Mean Suqred Logarithmic Error between the predicted and actual worldwide revenue predictions. This will prevent blockbuster revenue movies from overweighting the outcomes.

Project Design

Key to this project will be pre-processing the data. Several of the fields (actors, crew, director, writer, genre) contains comma-separated lists of names. This will make every entry in those columns nearly, or at least largely, unique. The particular parameters are likely to play a large part in determining the success of a particular film, so identifying commonalities will be crucial. I will follow the method outlined at https://datascience.stackexchange.com/questions/14847/multiple-categorical-values-for-a-single-feature-how-to-convert-them-to-binary-u to turn each item in each of those fields into individual columns in the dataset, one-hot encoding them for analysis.

Once the data is pre-processed and encoded I will then begin with some data exploration, starting with maximum and minimums for movie revenue, standard deviation, etc. to better understand the data being worked with. I will also split the sample data into shuffled train and test sets. Even though there is a separate test set used for this competition, I will leave it completely out of the training for two reasons:

1. It does not have any outcome data associated with it
2. There will be no opportunity to overfit to the test data with it not included

This assumes that there will be similarities in actors, directors, etc. across both the sample and test datasets for the project.

I will then evaluate the effectiveness of a few model configurations against each other to compare the efficiencies of the training and testing scores. This will help me in identifying an appropriate set of parameters to use for maximizing accuracy while minimizing issues of overfitting. For final training I will also use k-fold to generate multiple train/test sets, which will further reduce overfitting and artificially expanding the amount of data trained againse.

I anticipate use of a DecisionTreeRegressor and GridSearchCV for training. With these I will be able to fit a model to the data that can be used for further predictions against the test dataset in the competition.