**Movie success predictor**

**Problem statement**:

Predicting whether a movie will be financially or critically successful based on previous datapoints like cast, director, date of release, budget, genre, run time etc.

**Prospective Clients:**

This project will help movie production houses on deciding whether or not to produce a movie. It might help cast members to make a decision on whether to pick up a particular role or not.

**Dataset**:

Either the movielens 20M database

<https://grouplens.org/datasets/movielens/20m/>

or imdb5000 movie database

<https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>

Movielens data is much bigger in size. It has 27000 movies and additional fields like user defined tags and user ratings. So it is a better database for recommendation engines as well.

**Approach**:

Step 1: Scrub the data to extract pertinent information.

Step 2: Find connections between different factors

Step 3: Use the data for training the algorithm

Step 4: Use it for prediction/evaluate the model

**Patent Litigation Dataset**

**Problem statement**:

Finding correlation between patent litigation and success/failure of companies, growth trends of the industry in question etc. Popular perception is that industries that have a lot of ongoing litigation are the most competitive so should have the most rapid growth. Analysis of this data will help uncover the truth behind this consideration.

**Dataset**:

[https://www.kaggle.com/uspto/patent-litigations](https://www.kaggle.com/uspto/patent-litigations%20)

**Fake Review Detection**

**Problem statement**:

Travel planning and reservations have completely moved to a web based medium. People are relying heavily on the reviews left by fellow travelers. So detection of veracity of reviews is more important than ever.

We can try to find a relation between factors like travel website, hotel location, hotel brand and keywords in the description and the truthfulness of the reviews. Then we can use the trained algorithm to predict the veracity of future reviews.

**Dataset:**

<https://www.kaggle.com/rtatman/deceptive-opinion-spam-corpus>

The training dataset has 4 fields of information hotel brand, positive/negative opinion of the review, travel website and the actual comments. For each of the reviews the dataset has information on whether the review was truthful or false.