Project B2: KAGGLE-IMDb
Predicting movie revenue
Team members: Kristjan Korela, Norman Pirk, Joonas Püks
Github repo

Task 2 - Business understanding

Identifying business goals:

Background

There are very many movies published each year. They all have a different set of resources available - the budget, the location of filming, etc. Also, the movies represent a variety of genres. Another thing that is different is how much money the movie makes, and that varies a lot. In many cases the gross income of the movie is one of the main success criteria of the movie. Therefore it will be interesting to study how certain characteristics of a movie affect its income.

Business goals

Our main goal is to find if and how a movie's year of release, budget, country and genres affect the worldwide income of the movie. We can find most of these connections by analysing and processing data, but in order to reinforce the findings we plan to use machine learning to predict the incomes of movies according to the parameters mentioned above. If we can train an algorithm to predict the income accurately enough, we can use it to evaluate current trends, and further development could lead to predicting future trends in movies.

Business success criteria

Assessing the situation:

Inventory of resources

List of available resources:

- 3 students from the Introduction to Data Science course
- IMBDb movies extensive dataset from Kaggle
- Hardware: each student has their personal laptop
- Software: Jupyter Notebook, git

Requirements, assumptions, and constraints

The project will be completed by December 14. The completed project will consist of a 3-minute video about the project, a link to the source code and a poster slide describing the project. The data used in the project is a public dataset from Kaggle, therefore there are no legal constraints or obligations.

Risks and contingencies

N/A

Terminology

N/A

Costs and benefits

Since this project is a student project and not developed by professionals, we do not expect to make any meaningful breakthroughs. But we still want to notice meaningful connections and develop an algorithm that can predict income with a reasonable margin of error. The cost of the project is not relevant, since it is made for academic purposes. The benefit is mainly the learning experience of a data science project.

Defining data mining goals:

Data-mining goals

Visualizing the initial dataset in order to find interesting relations and patterns.
Training a model that predicts the worldwide gross income of a movie based on the year of release, budget and country.

Data-mining success criteria

The model will predict the worldwide gross income with accuracy at least 0.9 on the testing data extracted from our dataset.
We will get new insight into the movie industry from visualizing the data we have.

Task 3 - Data understanding

Gathering data

Data requirements

To train an algorithm, we need a movie's year of release (integer), gross worldwide income (integer) and genres (string). While these are the required ones, we may use more of the existing data for training.

Availability

We found the data from Kaggle. The data was mined from IMDb and it has about 85,000 movies with 22 properties. It doesn't seem that we need more data than that.

Selection criteria

We have one csv file where all the needed data is. Currently we plan to use the "year", "budget", "genre", "country" and "worlwide_gross_income" fields. Since there are many genres and countries, we plan to remove movies that are from countries that don't make many movies or are from very niche genres. We also need to remove movies that don't have the essential information needed to train the algorithm.

Describing data

The dataset we will be using for our project has 85855 data points and 22 features which are: 'imdb_title_id', 'title', 'original_title', 'year', 'date_published', 'genre', 'duration', 'country', 'language', 'director', 'writer', 'production_company', 'actors', 'description', 'avg_vote', 'votes', 'budget', 'usa_gross_income', 'worlwide_gross_income', 'metascore', 'reviews_from_users', 'reviews_from_critics'.
We still need to prepare and process the data for our task. For example, the genre field can have up to three genres. We need to extract those genres and convey them with a binary property. Also we will adjust all the incomes to inflation to make the incomes and possible analysis conclusions clearer.

Exploring data

The features are from three different data types:
float64 (metascore, reviews_from_users, reviews_from_critics and avg_vote);
int64 (duration and votes);
object ('imdb_title_id', 'title', 'original_title', 'year', 'date_published', 'genre', 'country', 'language', 'director', 'writer', 'production_company', 'actors', 'description', 'budget', 'usa_gross_income', 'worlwide_gross_income').
When it comes to the features that we are interested in (year, country, budget, genre and worlwide_gross_income) there are some quality problems with the data. For example, there are several years which only have a single movie published. In the value set for the feature budget, most of the units are $, but there are some with also different units, for example RUR and KRW. Both country and genre features have values as single items and also sets, for example Japan and {Iran, Germany, Canada} or Drama and {Drama, Family, Action}. When it comes to NaN values, the counts for those in the needed columns are: country - 64, worlwide_gross_income - 54839, budget - 62145, and 0 for genre and year.
In order to proceed with the analyses and modeling we need to clean and rearrange the data so that

- the feature values with very low representation will be removed,
- the features with values as sets are separated into new features with binary values,
- all rows with NaN values in the columns that we will use for modeling and analyses will be removed.

Data quality

The main quality issue with the data is the large amount of NaN values in the budget and worldwide_gross_income value sets. There are several datasets about movies available on Kaggle, so in order to mitigate our problem we can check if some of the datasets have the information missing in this dataset.

Task 4 - Project plan

Part 1: Preliminary data analysis and preparation (DEADLINE: 6. dec)

- Plot movie income
    - Contribution in hours: Kristjan: , Joonas: 1, Norman:
- Adjust incomes to inflation and add a corresponding column
    - Contribution in hours: Kristjan: , Joonas: 1, Norman:
- Plot the genre popularities by year (volume or percentage?)
    - Contribution in hours: Kristjan: , Joonas: 1, Norman:
- Remove unnecessary columns and NaN values from the data that will be used for machine learning, and replace categorical values with boolean values using one-hot encoding.
    - Contribution in hours: Kristjan: 3, Joonas: 5, Norman: 5.
- (If needed, find additional data for the analyses, because there are a lot of NaN values in the columns that are of interest to us)
    - Contribution in hours: Kristjan: 4, Joonas: 5, Norman: 5.

Part 2: Machine learning using regression(or classifier) (DEADLINE: 10 dec.)

- Train a variety of machine learning models, and choose the one that predicts best on the testing data.
    - Contribution in hours: Kristjan: 10, Joonas: 4, Norman: 4.

Part 3: Preparation of results (DEADLINE: 14. dec)

- Prepare the poster slide introducing the project idea, methods and results
    - Contribution in hours: Kristjan: 2, Joonas: 2, Norman: 10.
- Create and publish a video about the project
    - Contribution in hours: Kristjan: 1, Joonas: 1, Norman: 3.