9/12/2020

PROFITABILITY PREDICTION OF MOVIE PROJECTS

Data Understanding

Team 6
Oguzhan Akan
Darius Hooks
Jaymish Raju Patel
Jason Sabal
Bibinur Zhursinbek

California State University, Northridge
COMP 541 – Data Mining – F2020
Assignment 2

2.1. Collecting Initial Data

Initial Data Collection

The Movies Dataset (MD)

- Source: Kaggle
- Cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, vote averages.
- Entity-relationship feature
- Data format: CSV
- Problems encountered: Corrupt revenue information
- Resolution: Use The Movie Industry Dataset for revenue which have the correct data.

Movies Industry (MI)

- Source: Kaggle
- Budget and revenue information of the movies between 1986-2016.
- Single, tabular format data
- Data format: CSV
- Problems encountered: Missing revenue and/or budget information
- Resolution: Filter out the missing data points from the data.

2.2. Describing the Data

Data Description

Database	**Table Name**	**Records**	**Fields**
Movies Database (MD)	Metadata	45466	11
Movies Database (MD)	Keywords	46419	2
Movies Industry (MI)	Movies	6820	18

- The Movies Database have two tables: "metadata" and "keywords". The first table includes basic information about the movie while the keywords table consists of related keywords for each movie. They can be linked together using "id" column.
- Movie Industry dataset have one table: "movies". This table includes all the relevant information about the movies, including our target, profitability.

2.2. Describing the Data

Types of the Data and Datasets

Data

- **a mix of structured and semi-structured data**
- MI dataset is completely structured
- MD dataset includes data where the values are either structured or semi-structured, particularly in JSON format.

Datasets

- MI dataset is a single-file tabular dataset
- MD dataset is a relational database data.
- Join criteria between MI and MD: movies' titles

2.2. Describing the Data

Types of instances of the Datasets - Movies

field	data_format	is_descriptive	is_target	data_attribute	example
budget	int64	0	0	continuous	8000000
company	object	0	0	categorical -> nominal	Columbia Pictures Corporation
country	object	1	0	categorical -> nominal	USA
director	object	1	0	categorical -> nominal	Rob Reiner
genre	object	1	0	categorical -> nominal	Adventure
gross	int64	0	0	continuous	52287414
name	object	1	0	categorical -> nominal	Stand by Me
rating	object	0	0	categorical -> nominal	R
released	object	1	0	continuous -> date	1986-08-22 00:00:00
runtime	int64	1	0	continuous	89
score	float64	0	0	continuous	8.1

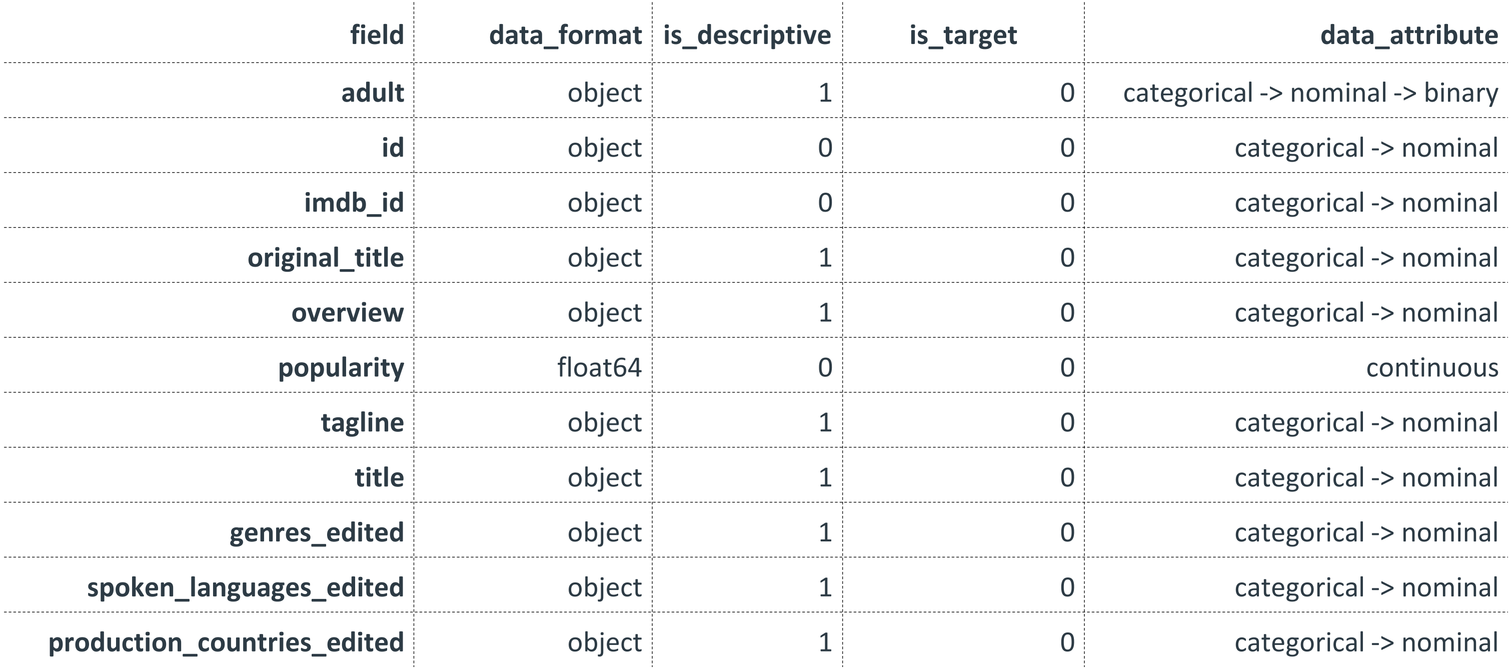
2.2. Describing the Data

Types of instances of the Datasets – Movies (Continued)

field	data_format	is_descriptive	is_target	data_attribute	example
star	object	0	0	categorical -> nominal	Wil Wheaton
votes	int64	0	0	discrete	299174
writer	object	1	0	categorical -> nominal	Stephen King
year	int64	1	0	categorical -> ordinal	1986
isprofit	int64	0	1	categorical -> nominal -> binary	1
profitability_ratio	float64	0	1	continuous -> ratio-scaled	5.53592675
profitability_ratio_bucket	int64	0	0	continuous -> ratio-scaled	550

2.2. Describing the Data

Types of instances of the Datasets - Metadata



field	data_format	is_descriptive	is_target	data_attribute
adult	object	1	0	categorical -> nominal -> binary
id	object	0	0	categorical -> nominal
imdb_id	object	0	0	categorical -> nominal
original_title	object	1	0	categorical -> nominal
overview	object	1	0	categorical -> nominal
popularity	float64	0	0	continuous
tagline	object	1	0	categorical -> nominal
title	object	1	0	categorical -> nominal
genres_edited	object	1	0	categorical -> nominal
spoken_languages_edited	object	1	0	categorical -> nominal
production_countries_edited	object	1	0	categorical -> nominal

2.2. Describing the Data

Types of instances of the Datasets - Metadata

field	data_format	is_descriptive	is_target	data_attribute
adult	object	1	0	categorical -> nominal -> binary
id	object	0	0	categorical -> nominal
imdb_id	object	0	0	categorical -> nominal
original_title	object	1	0	categorical -> nominal
overview	object	1	0	categorical -> nominal
popularity	float64	0	0	continuous
tagline	object	1	0	categorical -> nominal
title	object	1	0	categorical -> nominal
genres_edited	object	1	0	categorical -> nominal
spoken_languages_edited	object	1	0	categorical -> nominal
production_countries_edited	object	1	0	categorical -> nominal

2.2. Describing the Data

Types of instances of the Datasets - Keywords

field	data_format	is_descriptive	is_target	data_attribute
id	int64	0	0	categorical -> nominal
keywords_edited	object	1	0	categorical -> nominal

Thank you for your attention.