**EDA and Modeling using Rapid Miner Software**

**Table of Contents**

[**Problem Statement** 2](#_Toc46533162)

[**DS Pipeline** 2](#_Toc46533163)

[**Features and Data:** 3](#_Toc46533164)

[**Dataset:** 3](#_Toc46533165)

[**Preprocessing** 4](#_Toc46533166)

[**Exploratory Data Analysis (EDA)** 5](#_Toc46533167)

[**1.** **Reading CSV** 5](#_Toc46533168)

[**2.** **Select Attributes** 5](#_Toc46533169)

[**3.** **Set Roles** 6](#_Toc46533170)

[**4.** **Nominal to Numerical** 6](#_Toc46533171)

[**5.** **Split Data** 6](#_Toc46533172)

[**6.** **Replace Missing Values** 6](#_Toc46533173)

[**Modeling** 7](#_Toc46533174)

[**Results** 7](#_Toc46533175)

[**Apply Model** 7](#_Toc46533176)

[**Linear Regression** 9](#_Toc46533178)

[**Result History** 9](#_Toc46533180)

[**Explanation:** 10](#_Toc46533181)

# 

# **Problem Statement**

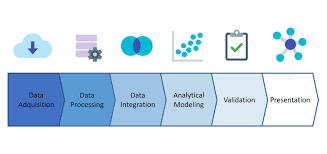
Netflix Movies and TV shows are released every year. Now a days there are far more number of Movies and shows that are released every day than ever before. In the present era it has become a challenge to know that either Movies or TV show will have highest number of audience worldwide. We have built a predictive model which will predict worldwide audience of Movies and TV shows in millions. We have trained our model using regression technique.

# **DS Pipeline**

A **data science pipeline** is the overall step by step process towards obtaining, cleaning, visualizing, modeling, and interpreting **data** within a business or group. **Data science pipelines** work for small **data** analysis.

Pipeline:

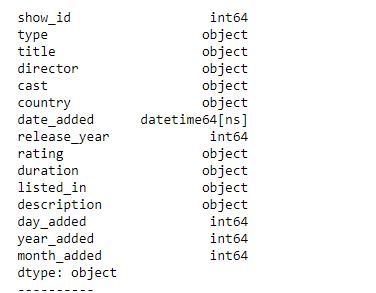
The following image shows how the dataset pipeline is followed for the process.



# **Features and Data:**

## **Dataset:**

We have selected Netflix Movies and TV Shows dataset which have record of different Movies and TV shows. The dataset has more than 10000 entries and it has the following attributes along with datatypes.

**Reason to choose this dataset:**

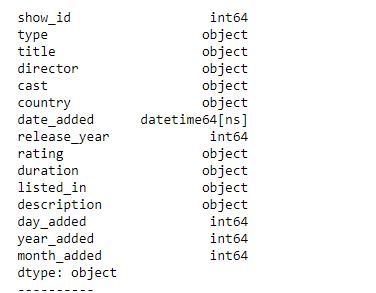
1. This was the most detailed dataset that we have searched so far aboutMovies and TV Shows. Plus we have very less missing values in this dataset
2. The data set was **.csv** file

# **Preprocessing**

1. Importing file into RapidMiner.

**Note:** as soon as we import data we can find out following things

1. Getting info about the dataset. The column names and their data types

****

1. Missing values present in the dataset

Replacing missing values:

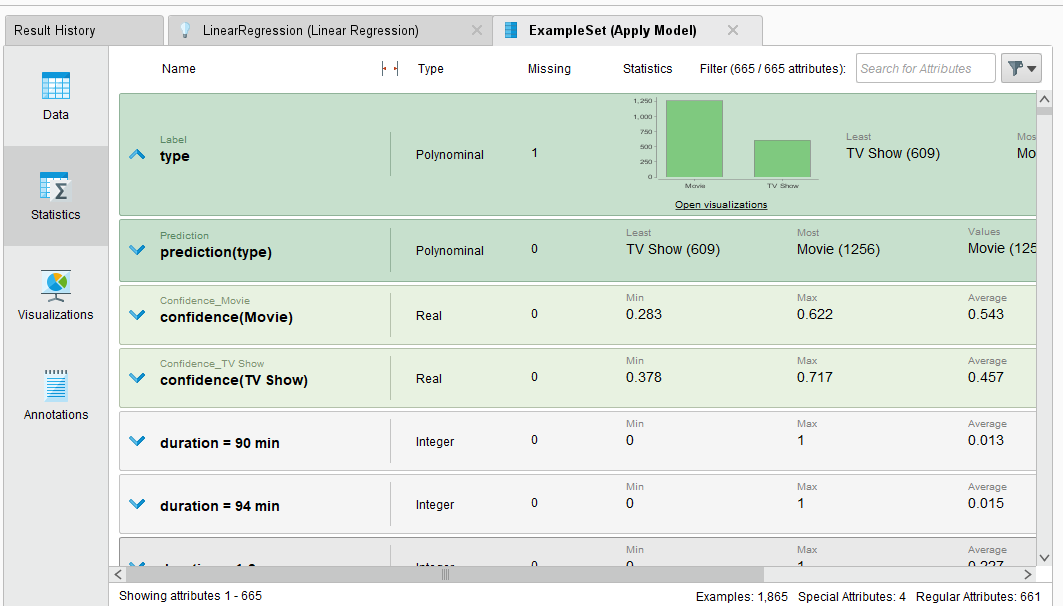
We replace missing values using replace missing values operator which replace values on the basis of average of that column.

**Results:**

After we run the process we can see imported data along with count of missing values, it’s statistics and visualization before the prediction.

# **Exploratory Data Analysis (EDA)**

After the preprocessing of the data exploratory data analysis is performed on the dataset through which we checked statistics and visualizations of selected data set.

****

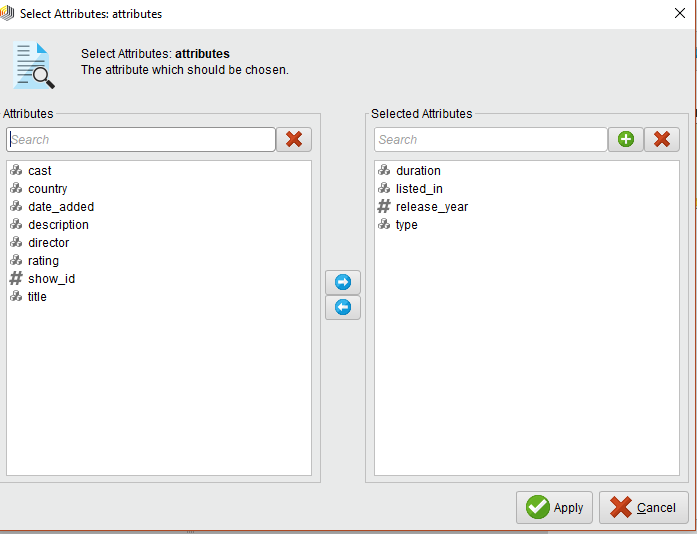
In the EDA folowing steps were followed:

## **Reading CSV**

First of all the CSV file was imported in the process. By using read CSV operator.

## **Select Attributes**

Secondly the attributes which were to be used for the modelling were selected.



## **Set Roles**

In this the role was selected for the prediction model. In our data set we selected the column “type” as role. And this column became dependent on all other attributes selected.

Type included two types:

* Movies
* TV Shows

## **Nominal to Numerical**

As we were going to apply regression model and this model only deals with numerical values we added nominal to numerical operator to solve the problem.

## **Split Data**

At this step we split our data in to training and testing data.

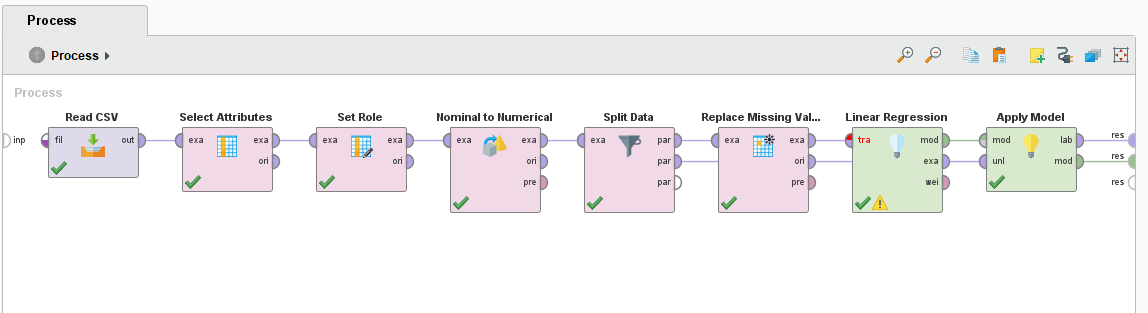
* 0.7 % training data
* 0.3% testing data

## **Replace Missing Values**

As our data contained some missing values, to solve that problem replace missing values operator was added.

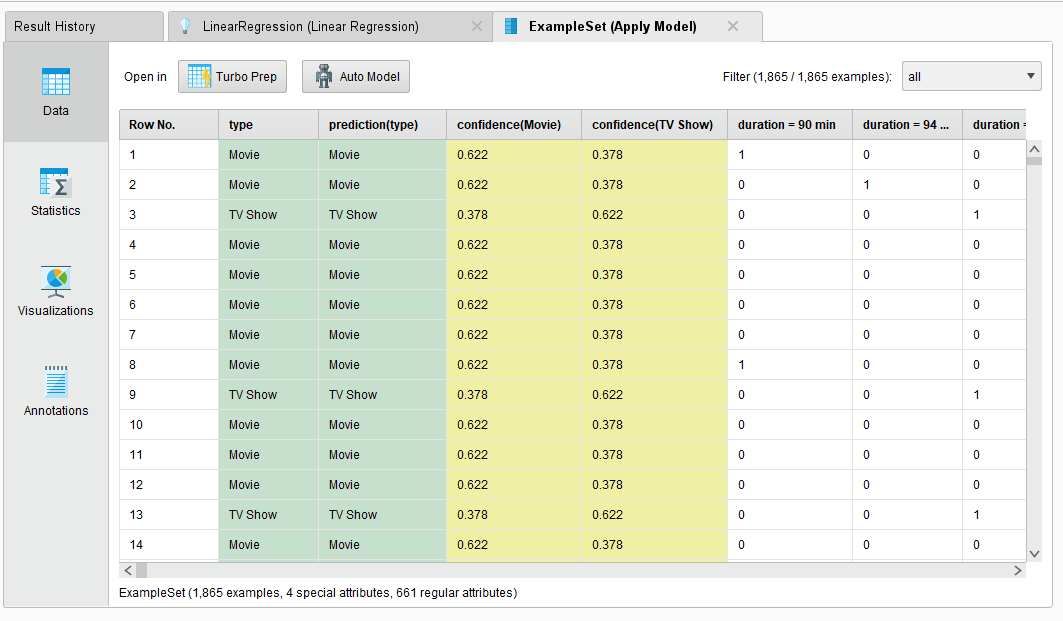
# **Modeling**

The regression model was used for the prediction purpose. The details are shown in the figure below:

As you can see there are two output connections. So we get of two types of outputs. The details of results are shown in the following figures.

# **Results**

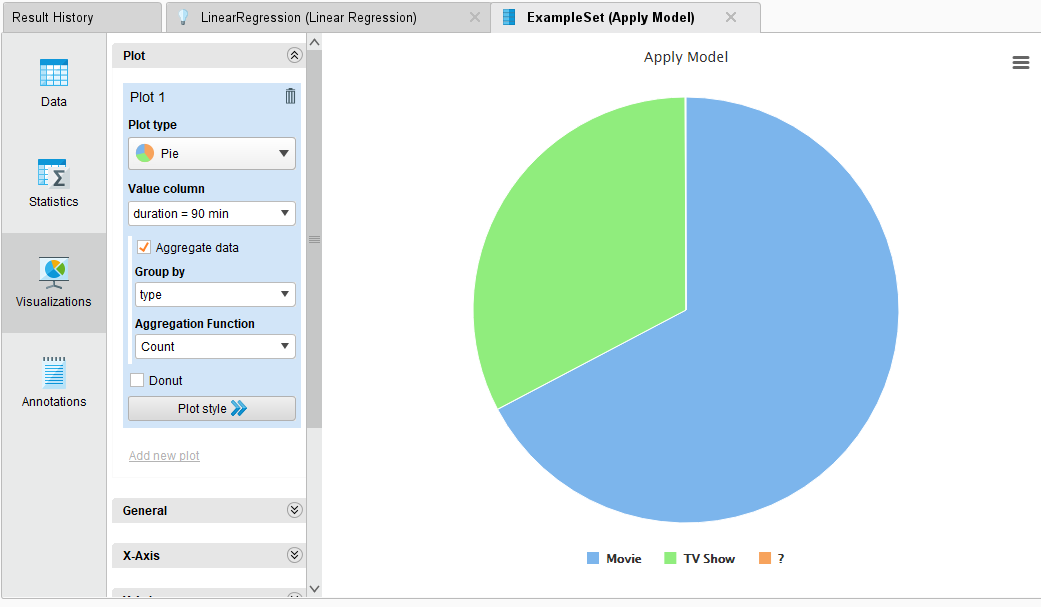
## **Apply Model**

**Data**

**Statistics**

# 

**Visualizations**

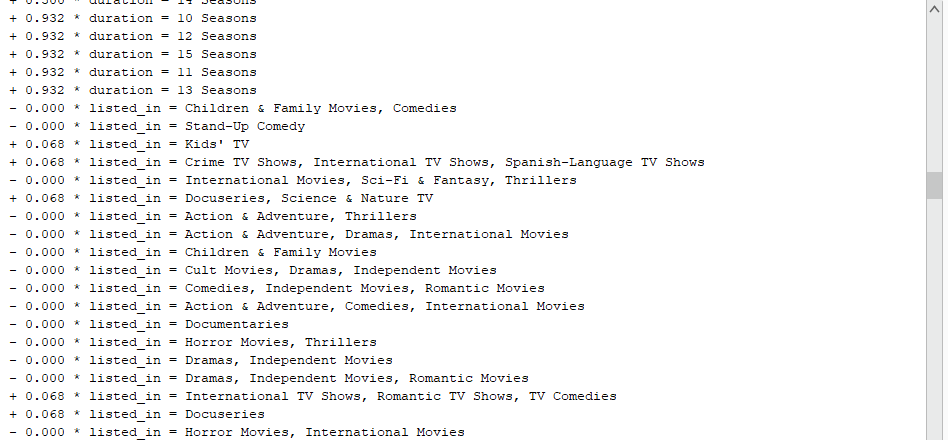
****

## **Linear Regression**

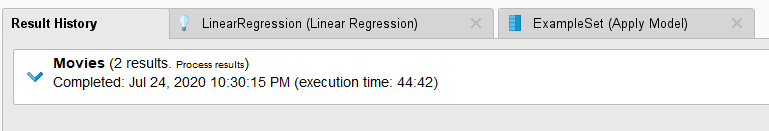
# 

Here we can get complete information about prediction using coefficient, p-Value and code.

**Description**



# **Result History**



# **Explanation:**

The above project is done using rapid miner studio version 9.7. Unlike python Rapid Miner provides interface and drag /drop options to make EDA easier.

We used following operators from operators menu:

* Read CSV
* Select Attribute
* Set Role
* Nominal to Numerical
* Split Data
* Replace Missing Values
* Linear Regression
* Apply Model

The data set was stored in Data repository.

We used two output connections and got above mention results.

Using the following attributes we predicted the popularity of type column in dataset.

* Listed\_in: It tells which kind of audience specific type has, either it is children, youth or old people.
* Release year: It tells the year of release of a particular Movie or Show.
* Duration: The time duration is mentioned in it which can affect the popularity of shows.

**Results**

The results shows that movies are going to be more popular then TV Shows in coming future.

We predicted it using coefficient, p-Value and code.

* **Coefficient**

If the coffiencent is positive it means that it is directly proportional to the role and if the coffiencent is negative it means that it is inversely proportional to the role.

* **p-Value**

Higher the p-value higher the co-relation.

* **Code**

In code column we contain asterisks. No asterisk means no co-relation. Less the numbers of asterisk more is the correlation.