Identifying your business goals

Background

Movie industry grow up every day and it is important to evaluate old and new movies. Movie producing is multi-part process and result of movie-making is peculiar product. It is complicated to understand is movie good or not but different approaches give opportunity for people to rate movies in order to describe for other in short form is movie should be watched or not? One of these approaches it is IMDb rating. IMDb - Internet Movie Database is an online database containing information and statistics about movies, TV shows and video games as well as actors, directors and other film industry professionals. Rating of IMDb it is numeric mark for different movies that people who have already watched movies give them.

Business goals

The main goal it is understand what makes movie high rated due to the available data on almost seven thousand films from IMDb. How to create really good movie that will have high score or maybe it is not possible at all? Second goal it is create program that will predict score of the current movie taking into consideration features of film (off course if we find that there is some relationship between the rating and this characteristic of the film).

Business success criteria

Our project will be done with success if we can prove that there is a certain algorithm that should be followed when making a film. Of course, we will be able to do this not only thanks to the numbers and percentages that we will receive, but also thanks to a deep analysis of some aspects of cinema.

- ## Assessing your situation
  - ### Inventory of resources

    All we need to make our project it is data about existing and rated films and technical knowledge of data science and machine learning, also some understanding of programing (python, Jyputer notebook). Data about movies will be taken from open internet resource that was mentioned – IMDb.

  - ### Requirements, assumptions, and constraints

    There is no specific requirements (except two students who will work on this project, one for data analyst and second for creating model for prediction). All data exist in free source so there is no problem with legal and security obligations. Our project must be completed by December 15.

  - ### Risks and contingencies

    There are no risks due to the cozy studying process provided by University of Tartu. If we could not continue develop our project at home, we have opportunity to do it at Delta center.

  - ### Terminology

    - Data science – it is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from noisy, structured and unstructured data.
    - Data analysis – it is a process of inspecting, cleaning, transforming, and modeling data.
    - Dataset – it is a collection of data.
    - Features of dataset - it is a collection of related feature classes that share a common coordinate system.
    - Correlation - it is a statistical measure that expresses the extent to which two variables are linearly related.
    - Correlation matrix - is a table showing correlation coefficients between variables.

- Cleaning dataset - data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
- One-hot-encoding – it is one method of converting data to prepare it for an algorithm and get a better prediction
- Machine learning – it is a field of inquiry devoted to understanding and building methods that 'learn', that is, methods that leverage data to improve performance on some set of tasks.
- Prediction model – it is a commonly used statistical technique to predict future behavior.
- Root-mean-square deviation (RMSD) or root-mean-square error (RMSE) - it is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.
- Random Forest – it is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.
- Overfitting - Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data.
- Accuracy – it is one of the most important performance evaluation metrics for a classification model. The mathematical formula for calculating the accuracy of a machine learning model is 1 – (Number of misclassified samples / Total number of samples).

Costs and benefits

Project created not for commercial purposes and required no costs.

- ## Defining your data-mining goals
    - ### Data-mining goals

      Firstly, we have to find out which movie is “the best” according to IMDb score list and “the worst”. Analyze what makes this movies high score (maybe some features are considerably important and others have no influence to score). Secondly, we have to use some data mining techniques and calculate correlations between specific features, maybe some features depend on others and it causes some consequences that affected to score. Finally, if any movie score consists of some relations, we can suppose that it is possible to create simplest model which will predict score of any exist movie and even predict score of not exist movie (for example we have some description of our movie that will be created and we want calculate score approximately).

    - ### Data-mining success criteria

      If we talk about the first goals, then it is worth emphasizing that, in general, finding interconnected values of the characteristics of a film will be a success, because films are multi-layered works that are difficult to describe through numbers. In our presentation, we will definitely indicate why the large values of some quantities, which should, in theory, bring glory to the film, may mean nothing in the case of some specific specifics. If there will be success among first two goals of our project, our model will have to predict score of the movie with RMSE lesser than 10% (I believe that this requirements are possible).