# MINI PROJECT CANVAS Title (preliminary): Movie revenues Group members: Samuli Siukonen, Simo Liimatta, Iiris Haapala Workshop # : 2 .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MOTIVATION 🎯*Which is the target group of our mini-project? Who is the end-user?* **Our project target group is movie producers and studios.**  *What are their objectives? What needs do we need to address with our work?***We aim to find optimal settings to produce a profitable movie.** *How will they benefit from this proposed solution?* **Make better decisions when producing a movie.** | DATA COLLECTION 🧩*Which data sources are we planning to use?***OMDb, TMDB, IMDB, grouplens.org, google**  *Mention database tables, API methods, websites to scrape, etc.***OMDb and TMDB have an API. Need to look more into scraping.** *Which is the data management plan?* **Possibly csv on hard drive. No confidential data is involved in the project.** | PREPROCESSING 🛠*What are the goals of the preprocessing pipeline?***Get clean data in easy and consistent format. Get good set of tags/keywords.** *Give some examples of data preprocessing steps.***Combining data from different sources. Removing missing values/unnecessary variables.** *What are some possible data cleaning/wrangling methods you’re planning to use?* **Combining data from different sources. Removing missing values/unnecessary variables.**  **Transform data.** *What are some possible data transformations that could be useful?* **Inflation correction. Taking percentage of revenue/budget.**  *Any feature engineering necessary?***No.** | EXPLORATORY DATA ANALYSIS (EDA) 🔎*Look at the data!  What steps are you planning to take towards exploring and understanding better the data you have?***plot all the variables against each other****see if there are any correlations. Look at the data.**  *What properties would be meaningful to summarize/visualize in this step?* **Revenue/review -ratio. Relation of the variables like above.** | VISUALIZATIONS 📊*List any meaningful visualizations you are planning to produce that will be useful to the end user?***Some graphs showing the relation between the revenue and budget, release time, leading cast, genre, review ratings, etc.** *Are you planning to produce any interactive visualizations?***Yes, if there’s time.** *If so, which types of interactivity might be useful to the end user?* **Maybe to filter out movies/keywords by different parameters.** |
| LEARNING TASK 🐭 (focus on problem definition) *Define the problem setting.* **Trying to predict revenue/ratigs from movie tags/key words. What kind of movies people want to see? What kind of movies are profitable to make?** *Is this supervised / unsupervised / other…?* **Supervised.** *Classification / regression / other…?* **Regression.** *What are we planning to learn? E.g. What is the target variable / learning outcome?***Profit(box office revenue)/reviews.**  *What variables are we using as input?* **Release time, movie genre, budget, rating, main cast(gender, age) + key words!** | LEARNING APPROACH 🪤 (focus on solution implementation) *Which ML/statistical methods seem more relevant for the defined problem setting and why?* **Linear regression? Seems like basic linear regression problem.** *Which evaluation metrics could be relevant?* Box office revenue. *Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?* **Test data should be collected randomly / evenly from the dataset.** |  | COMMUNICATION OF RESULTS 📢 *Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?* **A website might be the best way to show graphs.** *How do we communicate best our results to the predefined target group?* **Visual presentation and interactive website** *Short description of your interface/workflow (if applicable).* **Some interactive visualizations where you can filter data or show results according to different parameters.** | DATA PRIVACY AND ETHICAL CONSIDERATIONS 🔐 (if applicable)*Are there any fairness constraints that apply to our proposed pipeline?***No** *Is there a need to ask for consent during the data collection process?***Might need to ask for API-keys** *Is there a need for data pseudonymization/anonymization?***No need** *Any other privacy considerations that come to mind?* **No** |
|  | ADDED VALUE 🎁*Is there a possibility for added value from the data we’re planning to use?* *What is the added value?**How are predictions turned into added value for the end-user?* | **Movie studios and producers can make better decisions about what kind of movies to make.**  **Profit/movies that people want to see.**  **Let’s say we find a buch of tags/keywords or other variables that correlate positively with movies that have been very profitable. End-user can make decisions based on this information when making new movies, what actors make people watch more movies, etc.** |  | LEGENDWEEK 1: Data collection/preprocessing  WEEK 2: EDA & visualizations   WEEKS 3-4: Machine/deep learning  WEEK 5: Fairness & data privacy |