HW 10 (D14)

Karl Hendrik Bachmann, Emil Koemets, Jakob Univer

Introduction & Goals

Hi! We are a team of three: Karl Hendrik Bachmann, Emil Koemets and Jakob Univer and our project is “Analysis of Kickstarter projects information to predict possible outcome”. As the title says, our goal is to be able to predict the outcome of projects created in the kickstarter.com project starting community and therefore find the key elements of creating a perfect project.

Introduction

This project has not been done in cooperation with any companies, therefore there is no business goal in this project. However we do have our own goal set. Kickstarter.com is a website for startup companies who have an idea, but does not have enough money to put the idea into practice. On that Kickstarter website a company can create a project and then wish for a specific amount of money. Now the people who are visiting the website can pledge money to their desired projects and sometimes get even something in return. For example some projects have laid down some specific amounts of money, like “If you pledge us 5$, we will give you a special thanks in our project release video,” “For 10$ you can get an early version of our product.” Etc. But now here’s the catch. There’s always a deadline for every project in Kickstarter. If your project doesn’t get enough money till the deadline, compared to the amount you wished, your project gets cancelled and you don’t get any money. To get money, your goal must be reached before the deadline and this is where our data science project comes into play!

Terminology

In Kickstarter there are some terms that are not so self explanatory. For an example:

- Goal - Every project has settled an amount of money they want to get for the project. This is mostly referred to as a goal of their project.
- Pledge - A pledge is a promise to the project that if they reach their goal in time, I will give them the amount of money that I promised. It's important to be clear that it's just a promise. The money will be given only if they reach their goal. Otherwise your credit card won’t be charged.
- Backer - A person who has pledged.

- Reward - A backer can pledge their own chosen amount of money without getting anything in return, or choose a specific reward they want to get and pledge based on that. For example some rewards look like: “Pledge 5$ and you will receive a t-shirt with our logo”. Some rewards might have a specific amount and can’t be chosen over the amount. For example: “First 25 people who pledge 5000$ dollars will get to meet up with our crew.”

Our (data-mining) goals

Our goal is to find the key elements about making the perfect kickstarter project. Of course the most important thing is having a good project idea, but what if your kickstarter project gets cancelled just because you had a wrong amount of words in your project title, or you used a wrong currency for your financial goal? We want to be able to predict, whether the project succeeds or not, based on the title of the project, the category it’s in, currency, goal (how much money the project wants), how many backers the project already has (people who have pledged), the origin country of the project and how much money has already been pledged. And based on that we can also find out the keys of the perfect project.

Resources, assumptions & risks

Since we are using data that has been gathered by other people, we must assume that the data is correct. Also the resources we have are scarce. We must do all of the computing with our own computers. Talking about the risks there are a few big risks regarding this project.

Resources and resulting assumptions

As students we don’t have a budget to rent extra fast servers to do the fitting of classifiers, etc. We have to do the work on our computers instead. Also the data that we got from Kaggle (more info in “Gathering data” chapter), we expect it to be true, since checking the veracity of the data could be possible, but would be a very performance demanding process.

Risks

With this project comes a few big risks. One huge risk is that we don’t find any patterns (well this risk is basically included with all of data science projects). That means that the success of a kickstarter project is either random or only determined by the project itself. The latter is a measure we can not find in our dataset (since it’s impossible to measure the goodness of a project). Another huge risk is that we do find patterns, but the patterns we find are not

actually true patterns and therefore could be misleading our goals. Therefore it is important that we do different correlation tests to be sure that the conclusions we make from our project are true.

Task 3

Gathering data

We plan on using only one dataset, because it has information about over 370 000 projects from Kickstarter and we think that is enough to achieve our goal. This dataset can be freely accessed and downloaded from Kaggle. It is already in a computer-friendly csv format and no further processing had to be done to load that dataset with pandas. Most of the columns for this dataset are self-explanatory and those that aren't are documented on the Kaggle page. This dataset has duplicate columns of pledged amount in USD. One is taken directly from the Kickstarter project page and the other one is calculated with an external API(Fixer.io). Reading the discussion on Kaggle page we found that the direct conversion can be unreliable so we will not use that. Also there is a column of the pledged amount in local currency but that is not comparable so we will be using the one that is already converted to USD. Since we are trying to understand the success or failure of a project we won't be using rows with other states like cancelled. We did not face many difficulties gathering the data, because it was from Kaggle and in csv format.

Link to the described dataset: https://www.kaggle.com/kemical/kickstarter-projects

Describing data

After removing unwanted rows and columns described in the previous section, this dataset has 331 675 rows and 12 columns. As mentioned before the data already is in an easily readable format. This dataset has the needed columns and also since it contains all the projects from the launch of Kickstarter until the January of 2018 it has enough data for achieving our goals in our opinion. Here we will briefly describe each column.

Column	Description
ID	Unique identifier for the project

Name	Title of the project
Main category	Broader category for the project
Category	More specific category for the project
Currency	The currency used for the project given as three-letter ISO currency code
Deadline	Date when the project's success or failure is decided
Launched	Timestamp of the date and time when the project was launched
Backers	Number of persons who backed this project
Country	Country of the project given as two-letter ISO country code
USD pledged real	The amount of money pledged to project in USD
USD goal real	The amount of money needed to consider project successful in USD
State	The outcome of the project(success or failure)

Exploring data

Initial data analysis notebook can be found here: Notebook

Our dataset is fairly balanced outcome wise 40% of projects reaching its goal and 60% not. This should help us predict the outcome of a project more accurately since we do not have bias to one side. This dataset has 15 main categories, the most popular one being Film & Video and making up about 17% of the data. Also it has 159 more specific categories, the mode only making up 5% of the data. Most of the projects were launched in the US and used USD for currency, both of these make up about 80% of the data. All of the numerical values followed exponential distribution. Most projects in this dataset had 0 to 500 backers, had a goal of 0 to 100 000 US dollars and actually got pledged 0 to 50 000 US dollars. The launched and deadline dates followed both similar distributions when grouped by year, the most popular years being 2015 and 2016.

Verifying data quality

The dataset only had 3 missing values which were titles. Also there was one currency code which was incorrect since it had more than two-letters and 210 rows were affected by this issue. We did not find any other trivial data quality issues. This dataset has projects with either really small or high goals. These could be problematic since the low values mean that the creator did not really require any money or the project was a low effort attempt to get money. Also the same goes for the ones with really high goals since they probably had a really unrealistic idea if there were absolutely no backers. So for these projects we already know the outcome and this could skew our correlations between features.

Task 4

Project plan

Task	**Person**	**Hours**
Data exploring	Karl Hendrik Bachmann	1
	Emil Koemets	1
	Jakob Univer	1
Improving data quality	Karl Hendrik Bachmann	2
	Emil Koemets	2
	Jakob Univer	0
Feature engineering	Karl Hendrik Bachmann	4
	Emil Koemets	4
	Jakob Univer	4
Data visualisation	Karl Hendrik Bachmann	4
	Emil Koemets	4
	Jakob Univer	5
Modeling	Karl Hendrik Bachmann	13

	Emil Koemets	13
	Jakob Univer	13
Evaluation	Karl Hendrik Bachmann	6
	Emil Koemets	6
	Jakob Univer	7

Methods and Tools

We will mainly be using Jupyter Notebook for exploring data, feature engineering and modeling. We will also use Tableau for some of the data visualisation. We will use the CRISP-DM process as much as we can to make our workflow as efficient as possible. While modeling we will first try with simpler models and move on to more difficult ones to improve our modeling quality. The evaluation task will focus mainly on our project goals rather than evaluating how well the models work. In feature engineering task we will mainly try to create new features out of name, deadline and launch columns.