Analysis of Kickstarter projects information to predict possible outcome

Emil Koemets, Karl Hendrik Bachmann, Jakob Univer

INTRODUCTION

Our goal was 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.

OBJECTIVES

- Train a classifier so that it will predict the outcome of the project with a good accuracy.
- Find out the key elements of a good project
- Find ohter interesting facts and patterns in the dataset.

MATERIALS & METHODS

We used Jupyter Notebook for exploring data, feature engineering and modeling. We also used Tableau for some of the data visualisation. We used the CRISP-DM process as much as we can to make our workflow as efficient as possible.

DATA

We got the dataset from a Kaggle competition. It had information about over 370000 projects from Kickstarter. Our dataset is fairly balanced outcome wise 40% of projects reaching its goal and 60% not. 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.

PREPROCESSING

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 created new features from existing ones to make the dataset better for machine learning and looked for interesting patterns in the data.

MACHINE LEARNING

We tried three different approaches. Firstly we tried a usual random forest classifier on features category, main_category, currency, country and usd_goal_real. That approach gave an accuracy of 66.5 %. Next we tried making features of the most frequent words in project titles and train a random forest classifier with these features. This approach gave an accuracy of 68.7 %. Lastly we tried a LightLGB classifier on project titles which gave an accuracy of 70 % and a fairly good ROC curve.


RESULTS

Mean goal with confidence interval by outcome
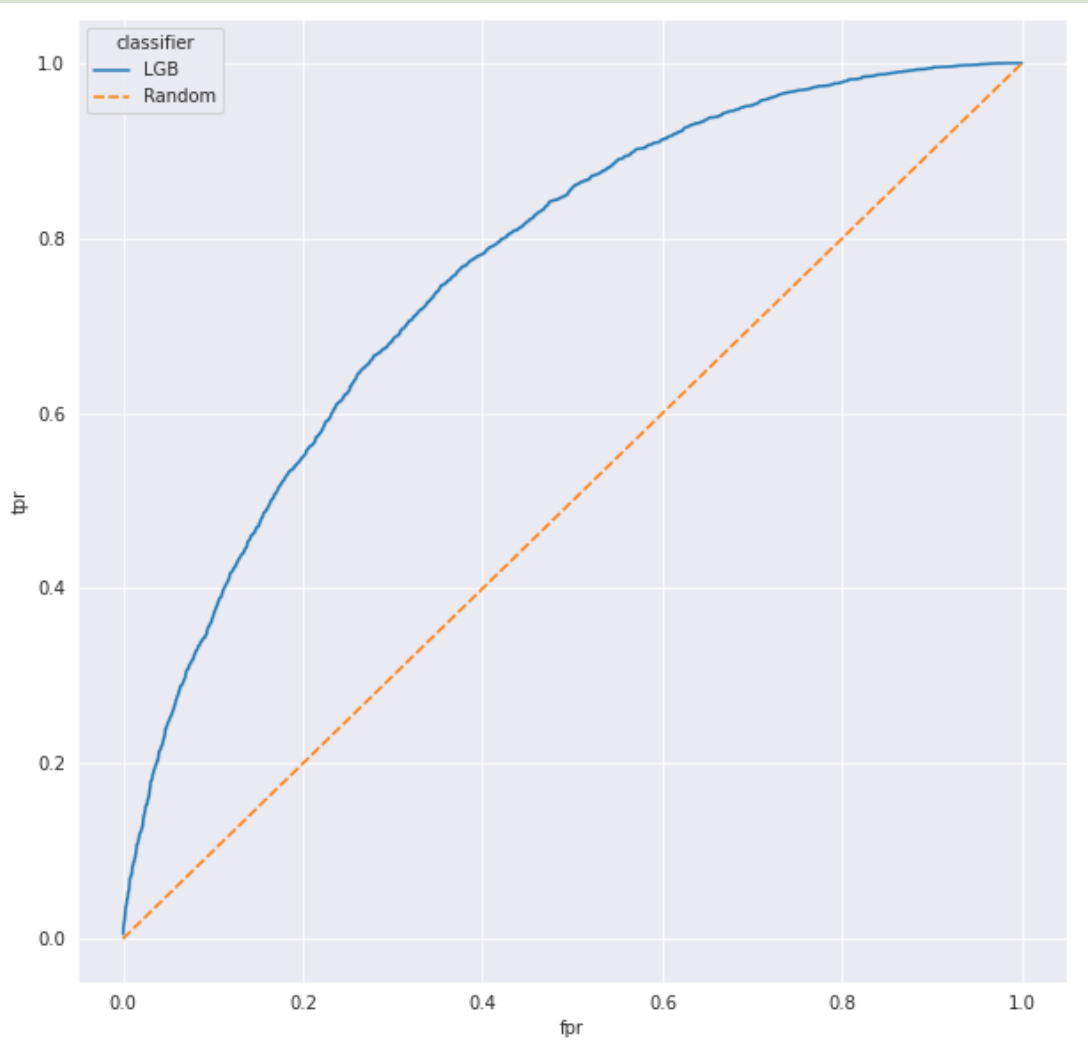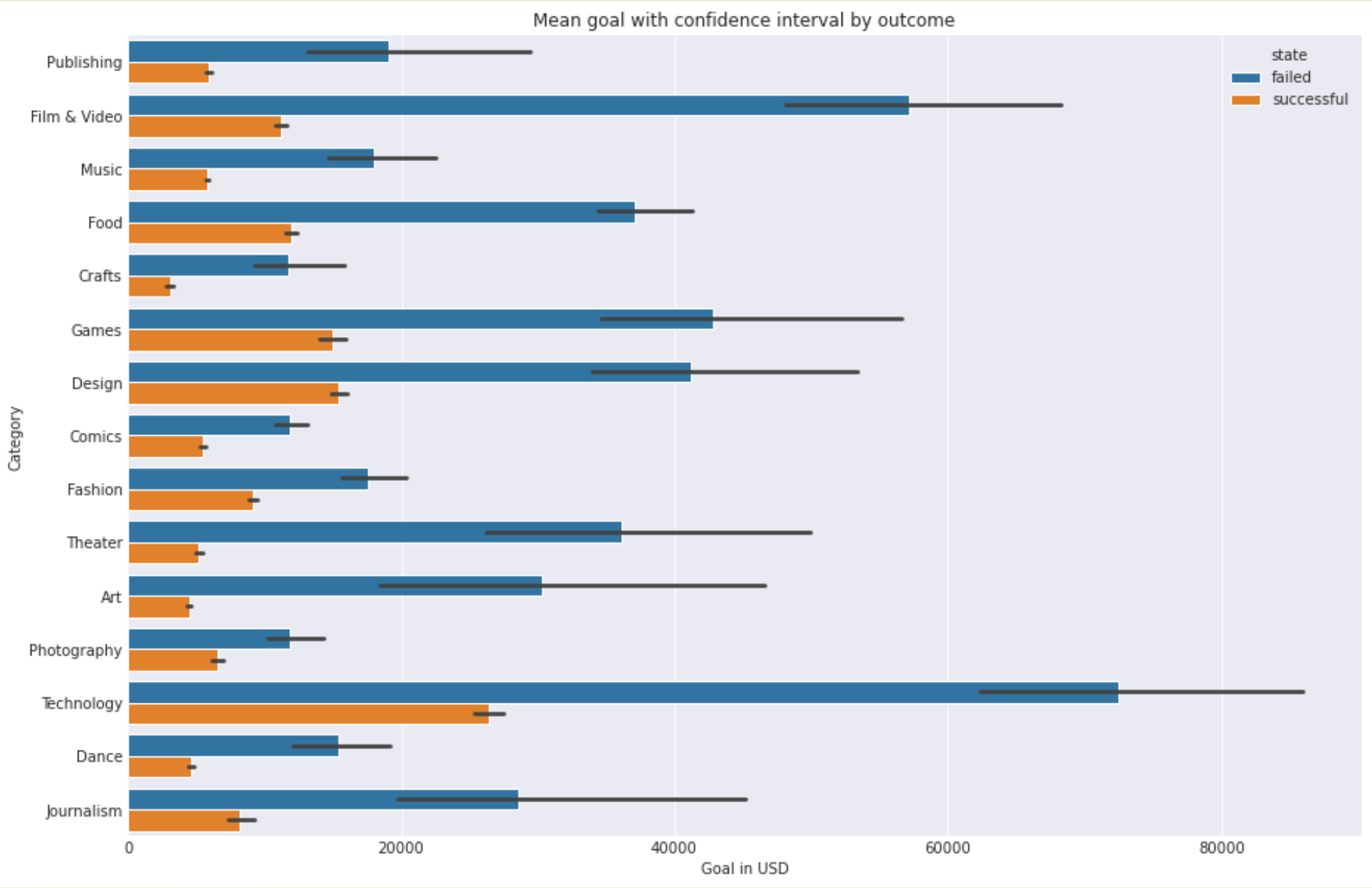
One of the most important features in all our classifiers was the goal amount. We found out that the mean goal amount for a successful project was on average about 3.5 times smaller than the amount for a failed one. 95% of the successful projects had a goal that was under 35 000 USD.

Feature importances
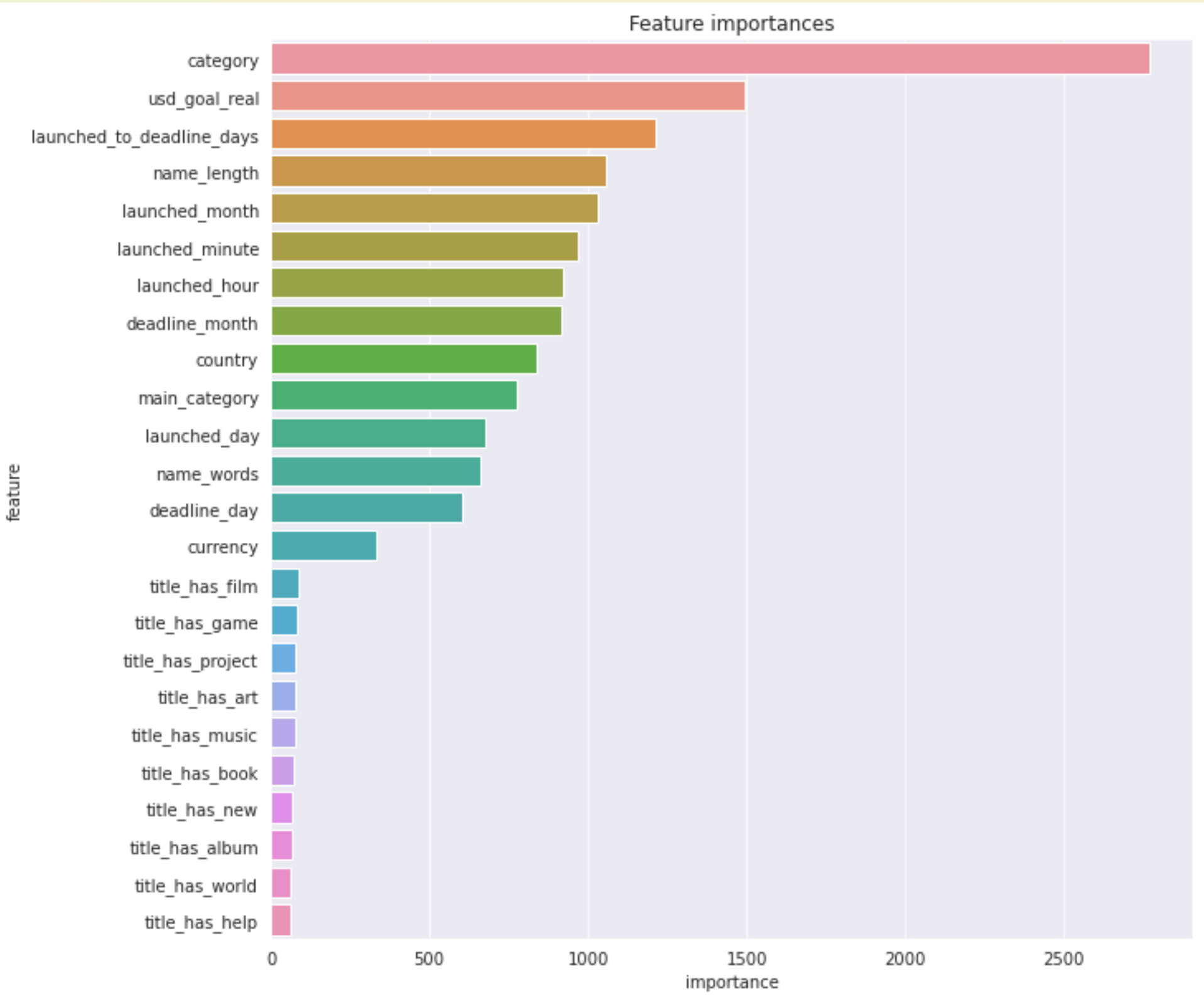
Another feature that had a high importance was the characters used in the title. We found that projects that use 40-60 characters that make up about 6-10 words in their titles have a better chance of being successful.
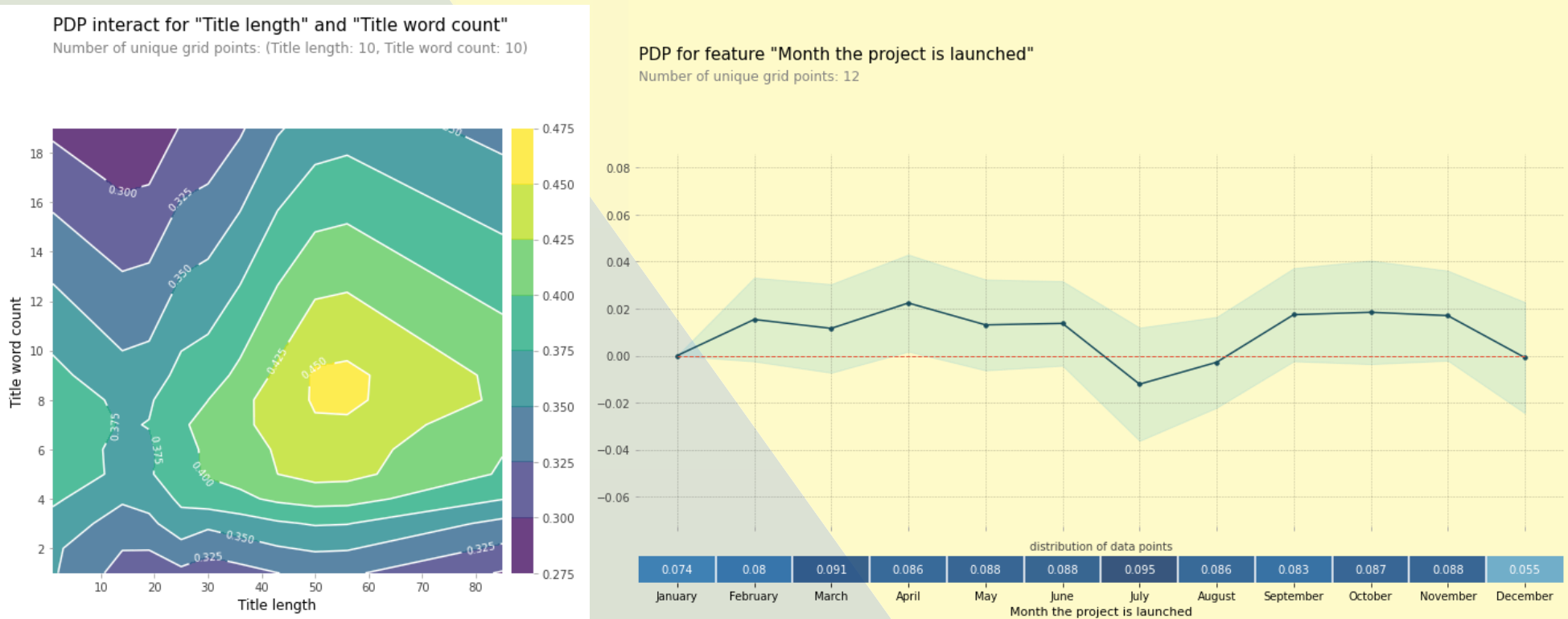
Month the project is launched also played an important role in the outcome. A project that is launched in March has a higher chance at succeeding. July was almost the only month that reduced the projects likelihood of being successful.

CONCLUSIONS

We did not reach our goal of being able to predict accurately whether a Kickstarter project will succeed or fail, but we managed to find out some key elements of a good project. We got our predictions up to 70% accuracy with project category and title as the only semantic features. We also found that on a dataset this big most classic classifiers are extremely time consuming to run.

REFERENCES

https://www.kaggle.com/kemical/kickstarter-projects
Source code: https://github.com/ekoemets/ids-project

CONTACT

jakob.univer@ut.ee
emil.koemets@ut.ee
karl.hendrik.bachmann@ut.ee