RELATÓRIO - ANÁLISES DE PROJETOS DO KICKSTARTER

Daniel do Carmo Granja de Castro

Guilherme Tamer Lotaif

Michel José Hanoch Vieira de Moraes

13.11.2018

SUMÁRIO

INTRODUÇÃO

Para o projeto final de Ciência dos Dados, fomos estimulados a escolher um dataframe qualquer, e trabalhar em cima dele para responder algumas perguntas. Para isso, trabalhamos com análise exploratória, e tivemos a escolha entre fazer um classificador, algum tipo de regressão, ou um projeto de clustering. Esse projeto fez com que os alunos pudessem trabalhar, agora num caso prático, tudo o que viram durante o semestre, nesta matéria.

TÉCNICAS

Para este projeto, algumas técnicas entre as que foram aceitas:

Clustering: Conjunto de técnicas de agrupamento automático de elementos por sua semelhança. A semelhança é um algoritmo escolhido de acordo com a necessidade da análise. Aos grupos formados por este processo damos o nome de cluster.

Classificador:

Regressão Logística: É uma técnica estatística que permite estimar o valor de uma variável categórica a partir de uma série de observações de várias outras variáveis independentes. Lembrando que as variáveis independentes devem ser explicativas ou binárias, sendo distribuídas conforme uma binomial.

IDEIA DE DATABASE

O nosso grupo escolheu trabalhar em cima de um database encontrada no “kaggle.com” com mais de 300 mil projetos do Kickstarter, uma plataforma de crowdfunding.

A pergunta a ser respondida durante este projeto foi “será que é possível prever se um projeto atingirá a sua meta apenas com as suas categorias iniciais?”.

VARIÁVEIS

Nome da variável	Descrição da variável
name	Nome do projeto
category	Categoria do projeto
main_category	Categoria específica do projeto
currency	Moeda do local onde o projeto está sendo produzido
deadline	Data de lançamento do projeto
goal	Valor desejado de investimento
launched	Data de início do projeto
pledge	doação total
state	Classificação de sucesso do projeto: - successful: projeto foi bem sucedido - canceled: projeto foi cancelado - failed: projeto falhou
country	País de origem do projeto
usd_pledge	Doação em dólares

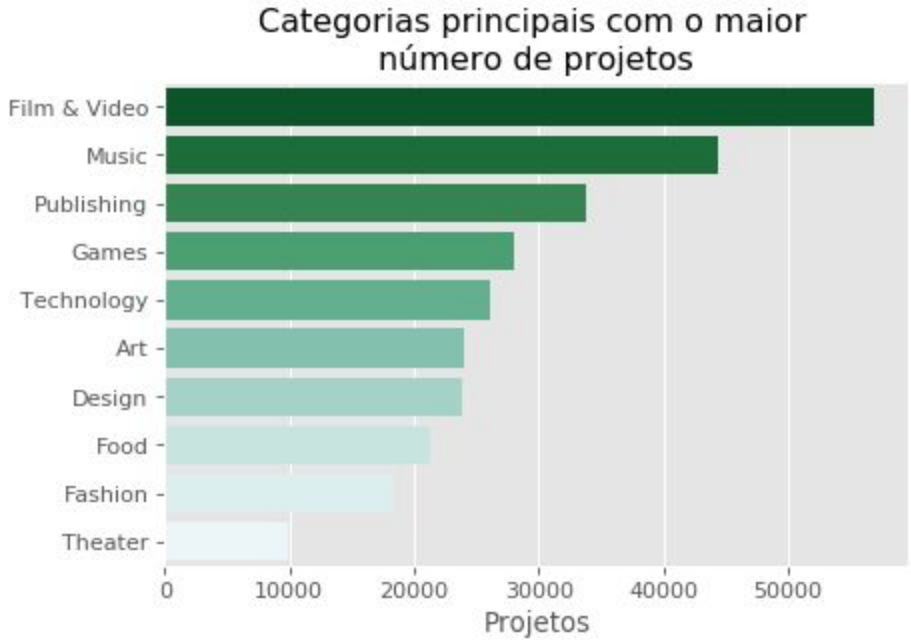
PROCEDIMENTO

Limpeza de Dados:

Começamos a trabalhar com esses dados fazendo uma limpeza geral, e criando um novo data frame usando o método get_dummies do One Hot Encoding para transformar colunas com variáveis qualitativas em variáveis quantitativas.

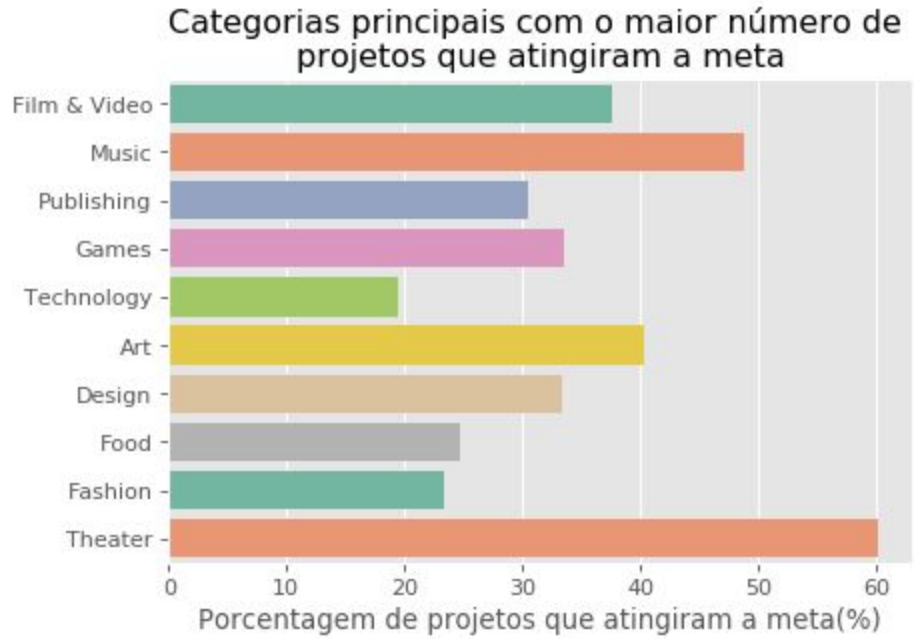
Análise Exploratória:

Nesta etapa trabalhamos com as variáveis do database. Primeiro fizemos a contagem do número de categorias e categorias principais, depois a contagem do número de projetos por categoria. Após descobrir as categorias com o maior número de projetos criamos um gráfico para as categorias e outro para as categorias principais por ordem de maior quantidade de projetos. Para criar os gráficos utilizamos os recursos das bibliotecas matplotlib e seaborn do python.



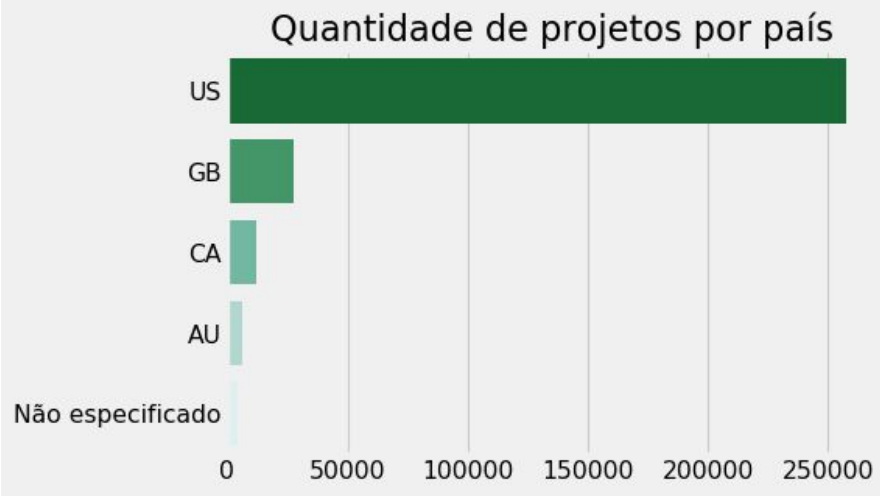
Com esses dados, descobrimos que a Categoria Principal com a maioria dos projetos é Film & Video.

Calculamos a porcentagem de sucesso por categoria principal e criamos um gráfico para visualizarmos esta taxa de sucesso.



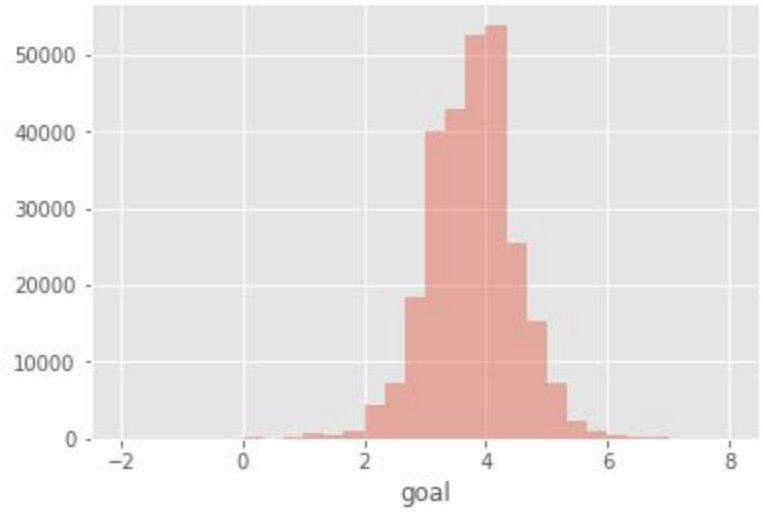
Pelo gráfico é evidente que teatro e música são as categorias com maior taxa de sucesso dentre as outras categorias.

A próxima etapa foi analisar os países, contamos o número de projetos por país e criamos um gráfico.



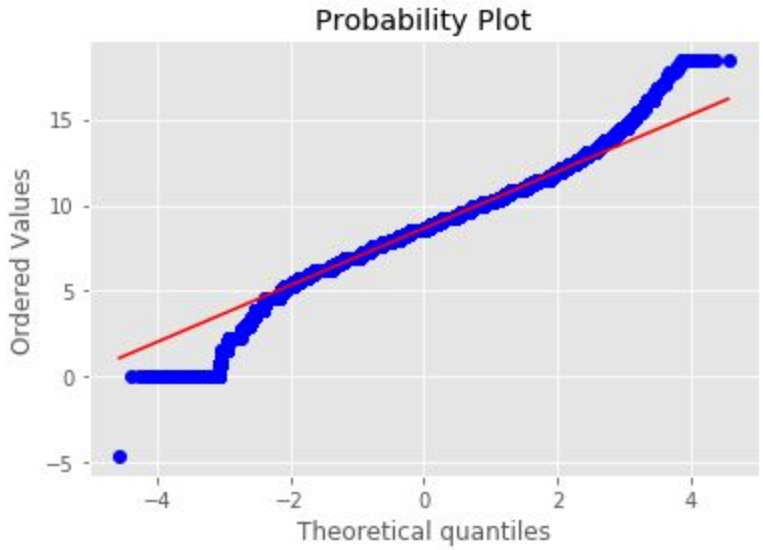
A grande maioria dos projetos são dos Estados Unidos.

E para finalizar criamos um gráfico para analisar a distribuição dos projetos conforme o seu valor de goal.



Percebe-se pelo gráfico que a maioria dos projetos necessitam de um valor de goal alto para que possam se concretizar.

Fizemos um probplot para conferir se esta distribuição se comporta como uma Gaussiana.



A partir do histograma ficamos com a impressão de que a distribuição se comporta como uma normal, porém ao analisar mais de perto vemos que não é.

Classificador DecisionTreeClassifier:

Depois de ter algumas hipóteses em mente, e de ter realizado toda a análise, e limpeza, dos dados. Resolvemos fazer um classificador para dizer se um projeto daria certo ou não.

Pesquisamos alguns tipos de classificadores diferentes e, no final, a decisão foi de usar um classificador chamado Decision Tree Classifier, com algoritmo criado por J. R. Quinlan, pesquisador de ciência dos dados, formado em Física e Computação pela Universidade de Sydney.

Esse classificador pega algumas “features”, e com base nelas tenta atingir um alvo, por meio de ramificações, que, calculando probabilidades, vão ficando cada vez menores, e mais detalhadas. Ou seja: ele checa se um determinado sujeito possui a “feature” hipotética “featureA”, partindo disso confere se também possui “featureB”, e assim por diante. Por cálculos de probabilidades, depois de uma base de treinamento, se torna capaz de dizer se determinado alvo é atingido ou não em uma “featureY”.

A maior diferença entre este classificador, e o classificador utilizado para o projeto 2, Naive Bayes é o fato de que além de analisar as categorias e suas independências, este classificador inclui possíveis dependências entre variáveis, sendo assim, um classificador mais completo para esta ocasião.

No caso deste projeto, separamos a coluna “state”, que indica a situação atual do projeto, em cinco colunas diferentes, e usamos apenas uma: “successful”, que era o “alvo” deste nosso classificador, para que no final ele pudesse classificar se, com base nas outras variáveis, descritas acima, o projeto atingiria este “alvo” ou não. Ou seja, ao definir um dataframe “X”, com todas as features que usamos para criar esta árvore, e uma coluna ”Y”, com a coluna alvo, e usar o algoritmo DecisionTreeClassifier, este programa é capaz de predizer se um projeto, que já definiu essas “features” iniciais, vai alcançar a meta, e portanto atingir o “state successful”.

Entre as maiores dificuldades encontrada pelo grupo durante este processo, está encontrar algumas falhas em colunas do dataframe, como troca de categorias, e strings em variáveis, teoricamente, numéricas, o que fez com que a acurácia mostrasse números completamente fora de realidade. Ao perceber isto, fizemos uma segunda limpeza de dados.

RESULTADOS

Tendo feito isso, usamos 2 terços da base de dados que foram separadas apenas para testes, para calcular a acurácia desse classificador, que antes de adicionar as variáveis: mês, ano, e meta estava próxima dos 67% de acerto, o que era proximo ao código do classificador RandomForest, que é bem preciso.

Depois de adicionar as últimas variáveis acima, a acurácia do classificador subiu para quase 100%.. Então preferimos retirá-las da análise para não termos um resultado equivocado, já que existia a possibilidade de ter mais falhas na base..

CONCLUSÃO

No final do projeto concluímos que com um simples código de classificador, podemos ter uma razoável precisão de se o projeto dará certo ou não com base nas variáveis antes descritas. E que se aprendermos a controlar melhor as variáveis, no nosso caso as datas, poderíamos ter um resultado bem mais preciso.

REFERÊNCIAS

1. https://www.saedsayad.com/decision_tree.htm
2. Referências de código no notebook do Jupyter.