Nanodegree em Engenheiro de Machine Learning

Aluno: Fernando de Lucca Siqueira

Predizendo sucesso de projetos do Kickstarter

Introdução e Proposta

Financiamento coletivo é uma forma de arrecadar fundos para um projeto através de investimentos comunitários. Ao invés de existir um banco ou empresa que financia o negócio, diversas pessoas financiam o projeto. Um dos sites pioneiros nessa prática na internet é o Kickstarter. Kickstarter ("chute inicial" em tradução livre é um site que conecta pequenas empresas ou indivíduos que querem lançar um produto mas não tem o capital suficiente para isso. Ao invés de tentar financiamento em bancos (que possuem juros altos), a pessoa pode recorrer ao site que gerencia o investimento de pessoas comuns no projeto.

Existem diversos tipos de projetos com diversos tipos de metas para serem implementados. Entretanto, investidores e criadores podem encontrar problemas no processo. Os investidores precisam de um direcionamento para saber se um projeto vai ter sucesso antes de investir seu dinheiro. Já os criadores precisam saber se seu projeto tem alguma chance de alcançar sucesso, caso contrário pode ser perda de tempo e dinheiro.

Descrição do problema

Para tentar auxiliar investidores e criadores, esse projeto se propõe predizer qual a chance de sucesso de um projeto no Kickstarter através de um classificador. Esse classificador irá predizer se um projeto com determinadas características vai ser financiado com sucesso ou não vai atingir a meta. Para evitar uma resposta binária, vamos utilizar a probabilidade do projeto se encaixar em cada classe do classificador.

Como o conjunto de dados para treino fornece a classificação dos projetos anteriores, esse problema se trata de um aprendizado supervisionado.

Conjunto de dados e entrada

Esse projeto irá usar uma base de dados com 378.661 projetos do Kickstarter (https://www.kaggle.com/kemical/kickstarter-projects). A base de dados contempla projetos de 2015 até 2018. A tabela abaixo descreve os dados do conjunto.

Campo	Tipo	Descrição
ID	int	identificador do projeto
name	text	nome do projeto
category	text*	categoria do projeto
main_category	text*	categoria principal
currency	text*	moeda aceita
deadline	date	data final do projeto
goal	float	meta para financiamento
launched	date	data de lançamento
pledged	float	total de financiamento alcançado
state	text*	estado final do projeto (sucesso, falha, cancelado, ativo, suspenso e indefinido)
backers	int	número de financiadores
country	text*	país de origem do projeto
usd_pledged	float	conversão do total financiado para dólar
usd_pledged_real	float	conversão do total financiado para dólar (atualizado)
usd_goal_real	float	conversão da meta para dólar (atualizado)

Os atributos com asterisco (*) serão considerados categóricos e transformados como colunas dummies. Será possível criar outros atributos como, por exemplo, tempo de campanha (deadline - goal) e fazer algumas análises com a quantidade de de financiamento levantado com o número de financiadores do projeto. O atributo **state** será a variável target do classificador, entretanto serão considerados apenas os valores de ‘sucesso’ e ‘falha’. A figura abaixo ilustra um exemplo dos dados.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged	usd_pledged_real	usd_goal_real
0	1000002330	The Songs of Adelaide & Abullah	Poetry	Publishing	GBP	2015-10-09	1000.0	2015-08-11 12:12:28	0.0	failed	0	GB	0.0	0.0	1533.95
1	1000003930	Greeting From Earth: ZGAC Arts Capsule For ET	Narrative Film	Film & Video	USD	2017-11-01	30000.0	2017-09-02 04:43:57	2421.0	failed	15	US	100.0	2421.0	30000.00
2	1000004038	Where is Hank?	Narrative Film	Film & Video	USD	2013-02-26	45000.0	2013-01-12 00:20:50	220.0	failed	3	US	220.0	220.0	45000.00
3	1000007540	ToshiCapital Rekordz Needs Help to Complete Album	Music	Music	USD	2012-04-16	5000.0	2012-03-17 03:24:11	1.0	failed	1	US	1.0	1.0	5000.00
4	1000011046	Community Film Project: The Art of Neighborhoo...	Film & Video	Film & Video	USD	2015-08-29	19500.0	2015-07-04 08:35:03	1283.0	canceled	14	US	1283.0	1283.0	19500.00

Métricas de Avaliação

A validação dos resultados será realizada em duas etapas. A primeira etapa será separar os dados em dados de treino e de teste. Os dados de treino serão usados para criar o modelo e o de teste para testar seu resultado. Os dados de teste serão avaliados a partir da métrica do F1-Score, que se baseia em uma média entre a **precisão** (precision) e **exatidão** (recall) dos resultados. A segunda etapa de validação será buscar outros projetos na plataforma que não estão no conjunto de dados, verificando qual foi a situação final e qual a predição do algoritmo.

Modelo de Referência

Os resultados obtidos com o modelo serão comparados com soluções propostas por usuários do kaggle (se possível) e com um classificador aleatório (DummyClassfier do sklearn pode ser uma opção).

Desing do Projeto

O projeto terá os seguintes passos principais: Análise exploratória de dados, criação de atributos, teste de modelos, validação dos resultados.

A análise exploratória de dados tem como objetivo entender a distribuição e organização dos dados, assim como identificar o comportamento e os atributos principais. Essa parte é essencial em qualquer projeto de data science, uma vez que ela permite entender os dados.

Na análise exploratória vamos confrontar a distribuição dos atributos, identificar as diferenças entre os projetos aprovados e reprovados, verificar o número de projetos pelos anos, tentar identificar sazonalidades, etc. Serão gerados gráficos com bibliotecas do matplotlib e seaborn para ilustrar os resultados.

Entendo como os dados se comportam é possível criar novos atributos derivados dos atributos iniciais ou que podem vir de outras fontes de dados. A criação desses atributos é resultado da etapa anterior, onde entendemos como os dados se comportam. Os novos atributos devem complementar o conjunto de dados. Um caso clássico de atributos a serem criados são as variáveis dummies para os dados categóricos.

Depois de identificar os atributos principais e os novos atributos é hora criar um modelo que consiga predizer os resultados. Esse problema é um problema de aprendizado supervisionado e classificação.

Seguindo o paradigma da metodologia ágil, inicialmente serão testados modelos simples, aumentando a complexidade e modificando os parâmetros de acordo com a necessidade. Para cada modelo será usada a técnica de GridSearch com cross-validation, identificando os parâmetros otimizados para predizer melhor o resultado.

Dentre os possíveis métodos para classificação podemos citar Logistic Regression, Decision Tree, Random Tree, Naive Bayes, SVC, Neural Networks, etc. Para agilizar o processo serão utilizadas as implementações destes métodos do framework scikit-learn (http://scikit-learn.org/).

Os modelos de predição com os parâmetros otimizados serão confrontados com os dados de teste para selecionar o melhor modelo. Posteriormente o mesmo modelo será testado com o segundo conjunto de teste, retirado do próprio site do Kickstarter para evitar possíveis sobreajuste com os dados originais.