Prediciendo el éxito de un proyecto en la era "crowdfunding"

Gustavo Augusto Mondragón Sánchez 136894
Francisco Martín Pérez López 136869

Aprendizaje máquina

Profesor: Carlos Fernando Esponda Darlington

Proyecto Final 2017

Introducción

Kickstarter es una empresa fundada en 2008 cuyo modelo de negocios se denomina "crowdfunding". La compañía tiene como medio de llegar a sus clientes un servicio web donde publica diversos proyectos "creativos" en el que cualquier persona con acceso a internet y una tarjeta bancaria válida puede patrocinar un al proyecto de su elección.

A partir de la fecha en que se publica la idea, los creadores del proyecto a fondear tienen un plazo definido en el cual recibirán contribuciones de patrocinadores a los que les haya gustado su proyecto. A cambio de las contribuciones, los creadores del proyecto, darán algo a cambio (desde una tarjeta de agradecimiento, hasta el producto finalizado o algo más.).

En caso de no se llegar a la meta en el periodo establecido el dinero se regresa a las personas que lo donaron. Es así como el éxito de un proyecto en Kickstarter consiste en recaudar el dinero propuesto en la meta y en el lapso de tiempo estipulado en la publicación del proyecto.

Hay diversas categorías en las que pueden pertenecer los proyectos: arte, cómics, danza, diseño, moda, cine y vídeo, comida, juegos, música, fotografía, publicaciones, tecnología y teatro por lo que en la página se puede encontrar, y apoyar, proyectos para los gustos más diversos.

En 2013, Kickstarter ganó "Best Overall Startup", sin embargo, actualmente la imagen de la compañía se ha visto perjudicada ya que se han dado varios casos en los que los donantes se sienten defraudados y debido a las políticas de Kickstarter no pueden reclamar, ya que solo advierte a los inversores que decidan a qué proyecto financiar de acuerdo a su criterio y que la empresa no es responsable y desconoce el fin último de los recurso obtenidos, pero también advierte a los dueños del proyecto que pueden ser sujetos a consecuencias legales de no cumplir con lo prometido en la publicación.
Por esta razón, es de suma importancia que los donadores puedan saber si el proyecto que van a apoyar tendrá éxito, pero también es muy importante para los administradores de nuevos proyectos que conozcan las características de los que han alcanzado la meta para que, con base en esto, creen mejores y más concretas propuestas comerciales y generen mayores probabilidades de éxito.

Vivimos en una era de cambios rápidos, todos los días en alguna parte del mundo nace una gran idea y muchas se convierten en Startups incluso algunas se consolidan como grandes compañías que cambian nuestra forma de ver el mundo como Facebook o Tesla motors.

Si bien la información que el proyecto provee puede ser usada por los patrocinadores el objetivo principal es el segundo, generar una herramienta para la toma de decisiones y para el modelado del éxito de un proyecto que permita, en nuestra nueva era, con tantas variables que tener en cuenta, crear una propuesta con mayores probabilidades de éxito.

Metodología

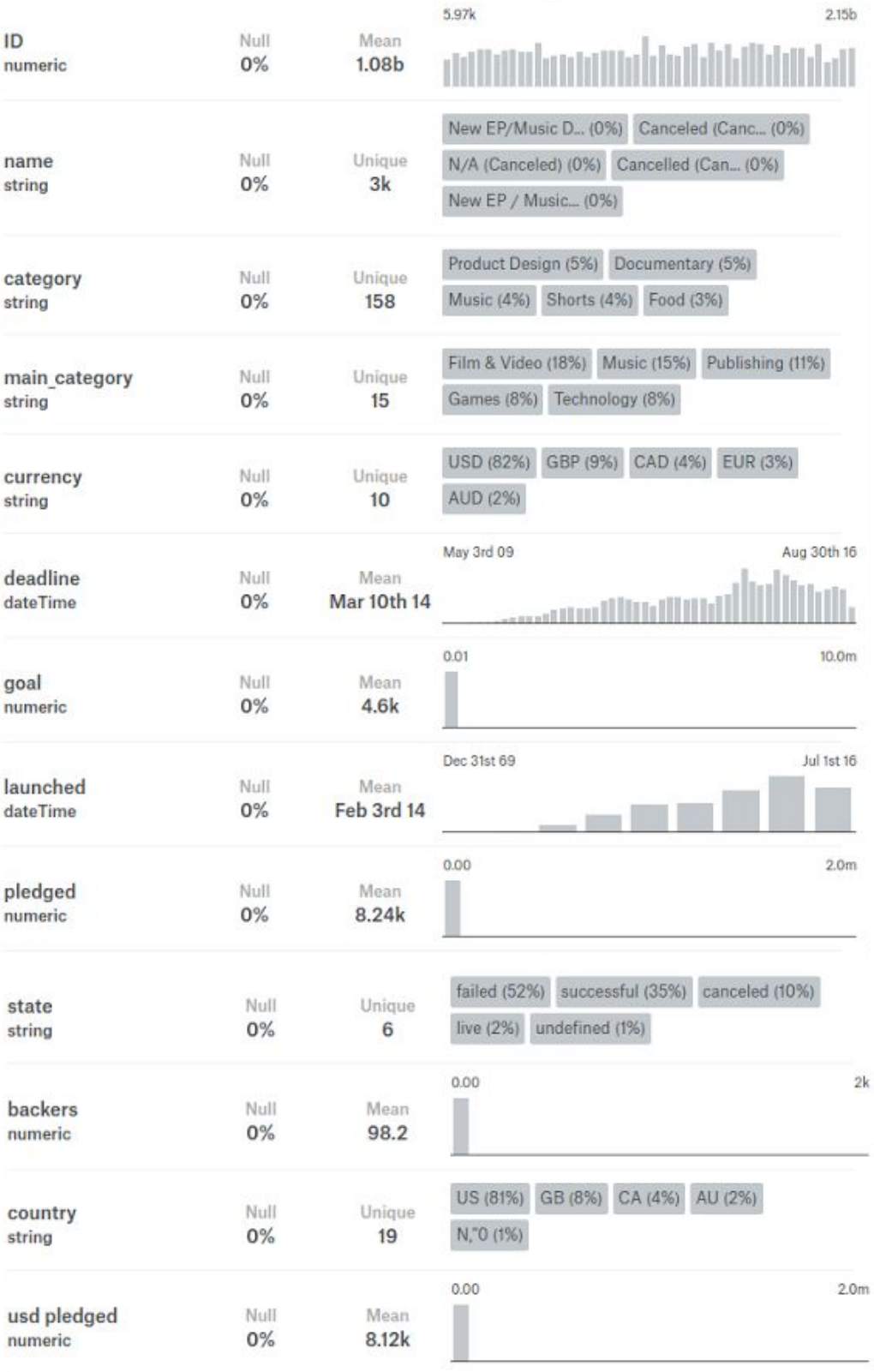
Librerías utilizadas

Se utilizó pandas para cargar los datos, numpy para el fácil manejo de las variables, matplotlib y seaborn para las visualizaciones gráficas y sklearn como librería con modelos de machine learning.

Análisis exploratorio de los datos
Existen 300,627 proyectos registrados con 13 atributos:

Variable	Tipo	Descripción
ID	Numeric	Identificador único del proyecto
Name	String	Nombre del proyecto
Category	String	Subcategoría a la que pertenece el proyecto
Main_category	String	Categoría principal
Currency	String	Moneda en la que se fondea el proyecto
Deadline	DateTime	Fecha límite para contribuciones
Goal	Numeric	Monto requerido para alcanzar la meta del proyecto.
Launched	DateTime	Fecha en la que se liberó el proyecto en la página
Pledged	Numeric	Monto recaudado
State	String	Estado en el que se encuentra el proyecto
Backers	Numeric	Número de patrocinadores
Country	String	País del creador del proyecto
Uds_pledged	Numeric	Monto recaudado en dólares

Resumen preliminar: [1]



[1] Fuente: https://www.kaggle.com/kemical/kickstarter-projects/data

Limpieza de los datos

Además de la limpieza se crearon nuevas columnas que fueron incrementando la exactitud (las predicciones pasaron del 65-70% al 85-92%).
Las columnas creadas son:

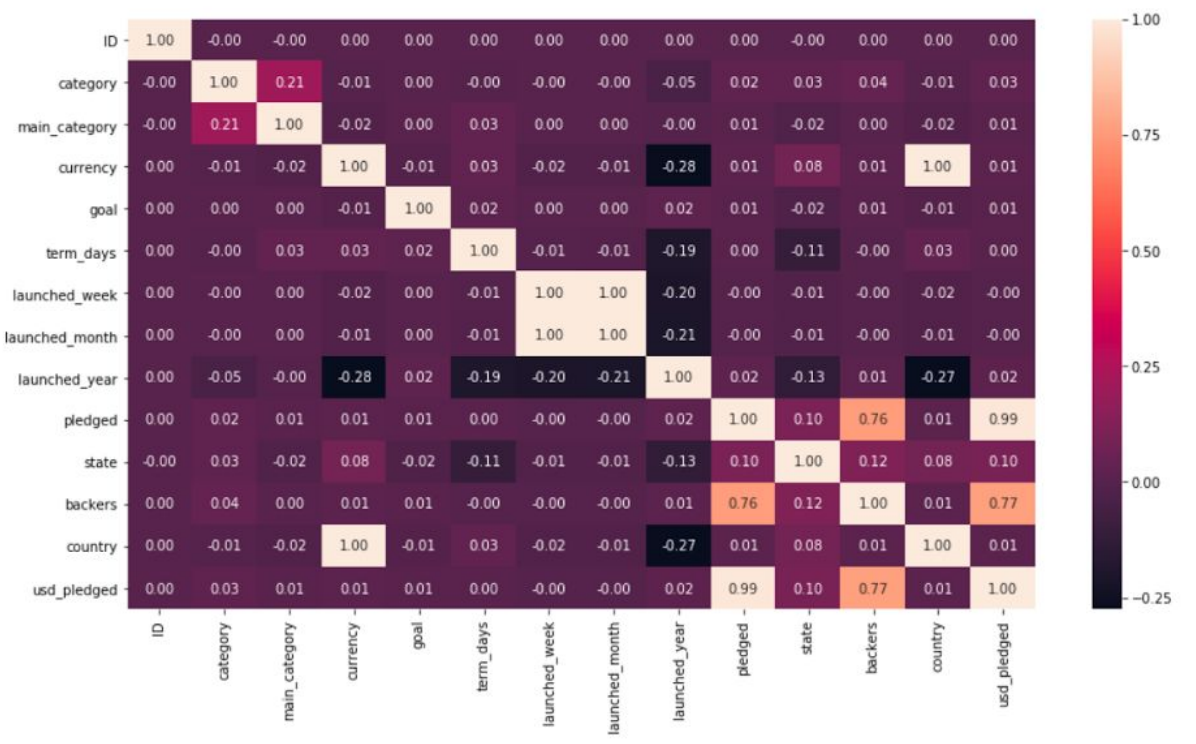
- term_days: la diferencia entre la fecha de término y la de inicio del proyecto.
- launched_week: semana del año en que se lanzó el proyecto.
- launched_month: mes del año en que se lanzó el proyecto.
- launched_year: año en que se lanzó el proyecto.

Se eliminaron los registros de 1970 ya que debido a la lejanía temporal (casi 50 años) y a que todos tienen un estado indefinido, en este momento, no aportan ninguna información útil a la predicción.
Se rellenan con “no_name” los únicos dos proyectos sin nombre
Se eliminan los registros de los proyectos que siguen activos o están indefinidos ya que no aportan información para nuestros algoritmos supervisados.

En nuestro análisis del proyecto definimos como un proyecto exitoso el que contenga como estado la etiqueta *"successful"* mientras que los proyectos fracasados son los que contengan como estado las etiquetas *"canceled", "failed", "suspended"*.

Se realizó una conversión de las demás variables categóricas (strings) a códigos numéricos.

Ya que los datos están en formato numérico buscamos relaciones entre las variables:



Con base en lo observado en la matriz de covarianza se decidió utilizar las variables siguientes:
category, main_category, currency, goal, term_days, launched_week, launched_month, launched_year, pledged, backers, country, usd_pledged.

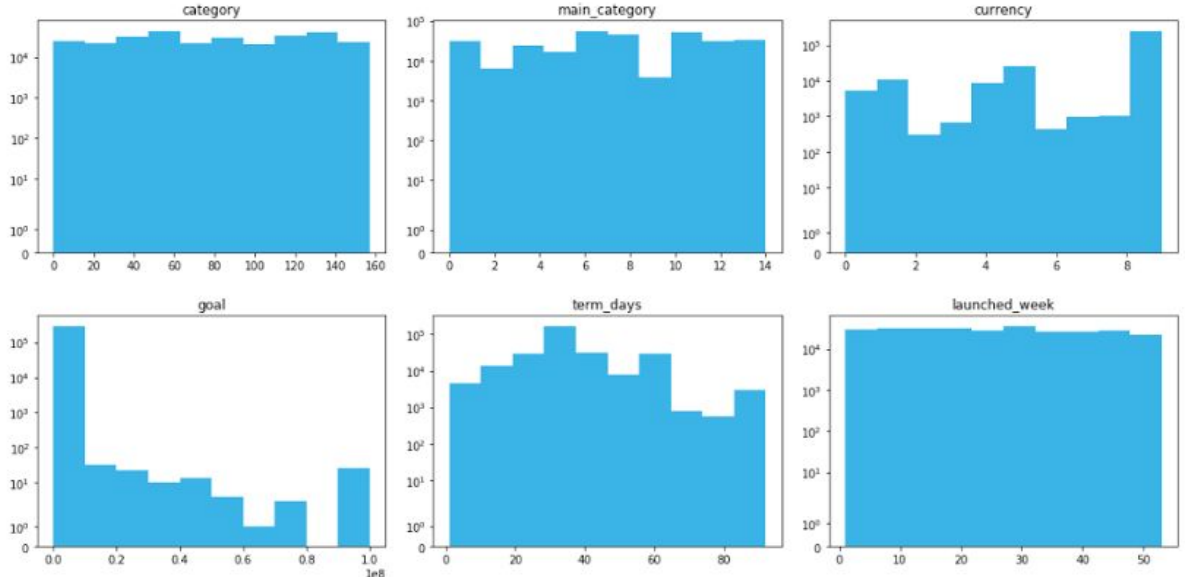
Análisis de las distribuciones de cada variable que se usará:

Variable objetivo: "State"

count	292132.000000
mean	0.359618
std	0.479889
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

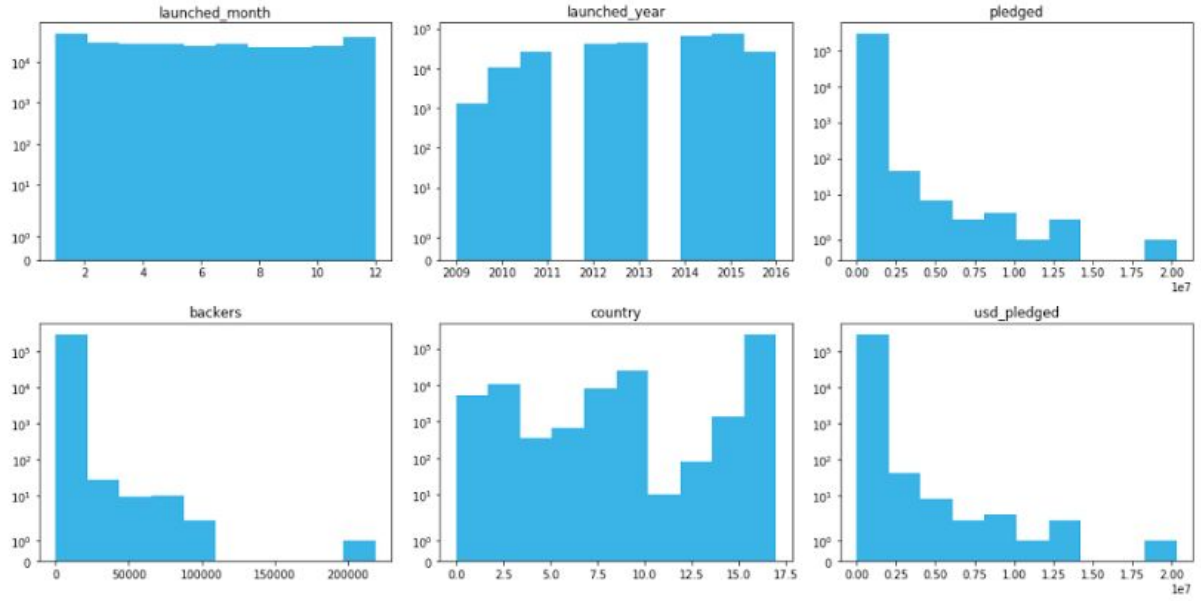
Variables predictoras (ejes logarítmicos):

	category	main_category	currency	goal	term_days	launched_week
count	292132.000000	292132.000000	292132.000000	2.921320e+05	292132.000000	292132.000000
mean	80.582295	7.530880	8.022664	4.656734e+04	34.491624	25.739816
std	44.168826	3.890814	2.263081	1.137343e+06	13.094158	14.483061
min	0.000000	0.000000	0.000000	1.000000e-02	1.000000	1.000000
25%	40.000000	5.000000	9.000000	2.000000e+03	30.000000	13.000000
50%	86.000000	7.000000	9.000000	5.000000e+03	30.000000	25.000000
75%	121.000000	10.000000	9.000000	1.500000e+04	38.227778	38.000000
max	157.000000	14.000000	9.000000	1.000000e+08	92.000000	53.000000



De las anteriores gráficas se puede concluir que los proyectos están, muy usualmente, patrocinados en dólares, que el monto objetivo, en los primeros tres cuartiles, es menor a 15,000 dólares. Los proyectos no suelen durar más de 90 días para recaudar el objetivo y en promedio se marcan una meta de un mes.

	launched_month	launched_year	pledged	backers	country	usd_pledged
count	292132.000000	292132.000000	2.921320e+05	292132.000000	292132.000000	2.921320e+05
mean	6.240751	2013.563742	8.316724e+03	99.810955	15.287644	8.162637e+03
std	3.332632	1.624536	8.778767e+04	901.431784	3.991360	8.714291e+04
min	1.000000	2009.000000	0.000000e+00	0.000000	0.000000	0.000000e+00
25%	3.000000	2012.000000	3.000000e+01	2.000000	17.000000	3.000000e+01
50%	6.000000	2014.000000	6.080000e+02	12.000000	17.000000	6.150000e+02
75%	9.000000	2015.000000	3.851000e+03	55.000000	17.000000	3.866000e+03
max	12.000000	2016.000000	2.033899e+07	219382.000000	17.000000	2.033899e+07



Los proyectos suelen ser lanzados en la página más veces en enero y en diciembre. Casi todos los proyectos tienen cero dólares acumulados, lo que ayudará a tener una buena base para predecir los proyectos que aún no han sido lanzados.

Transformación de variables

Se realizó una normalización de los datos, ya que mientras algunos están en categorías del 1 al 5, otros están en el orden de cientos o miles. Para que no influya la magnitud de los valores en las predicciones se llevan a la misma escala.

Separación del set de datos:
Después de todo el procedimiento anterior en el que se perdieron algunas tuplas quedaron 292,132 proyectos para hacer la predicción.
Se realizó una separación de los datos 80-20, quedando 234,031 proyectos para entrenar al modelo y 58,101 proyectos para hacer la validación de los modelos.

Creación de modelos

Modelo base (random)

El modelo base es una generación de números aleatorios, es decir, simplemente si la clasificación se dejara a la suerte.

Naive Bayes

Usualmente el modelo más básico es naive bayes, se usó la versión GaussianNB de sklearn. Lo tomamos como primer métrica de desempeño de un modelo inteligente. Al basarse en una predicción de las probabilidades calculadas de los datos no hay parámetros con los que se pueda mejorar el desempeño.

Decision tree

Se eligió una profundidad del árbol de 31 por ser la mínima que genera la mejor precisión en un rango de 1 a 100 niveles. Los parámetros específicos utilizados fueron los siguientes:

- criterion='gini'
- splitter='best'
- max_depth=31
- random_state=0

Random Forest

El número de estimadores es tres ya que es el valor mínimo que generaba una buena predicción, con 100 árboles el peso computacional no genera un gran incremento de la precisión. Los parámetros específicos utilizados fueron los siguientes:

- n_estimators=3
- criterion='gini'
- max_depth=31
- bootstrap=True
- random_state=0

Support Vector Machine

El único parámetro que incrementó la precisión fue el parámetro de penalización “C”, se probaron valores entre 1 y 30 resultando 10 como mejor parámetro. Los parámetros específicos utilizados fueron los siguientes:

- penalty=’l2’ (regularización l2 es la estandar para SVM)
- loss=’squared_hinge’: es una función de pérdida usada para maximizar el margen de clasificación, naturalmente es usada en SVM.

  La ecuación que describe a esta métrica es la siguiente: $\iota(y) = max(0, 1 - t \cdot y)$

  donde t es la salida del algoritmo y “y” la variable objetivo. Para optimizar la función se puede usar el gradiente (aunque no sea derivable se puede usar una modificación de la derivada con respecto a los parámetros del modelo “w”), la optimización de Rennie o la cuadrática. En este caso se utilizó la versión del optimizador cuadrático ya que presentó resultados que convergen en menor tiempo.
- C=10
- multi_class=’ovr’
- random_state=0
- max_iter=1000

Neural net con una capa oculta de 12 neuronas:

Se probaron las cuatro funciones de activación "identity", "logistic", "tanh", "relu" de las cuales relu fue la de mejor desempeño, mientras en la primera red alpha de 0.1 dio buenos resultados en la segunda fue 0.01 una mejor elección. Los parámetros específicos utilizados fueron los siguientes:

En ambas redes se probaron tres optimizadores (Descenso por gradiente estocástico, optimizador de Adam y L-BFGS) pero se decidió utilizar el optimizador L-BFGS ya que presentó los mejores resultados (más rápida convergencia y mejor precisión de predicción).

L-BFGS es un método de optimización quasi-Newton para funciones complejas. La idea principal es hacer uso “limitado” de la memoria. El cálculo del optimizador se hace con calculo matricial, sin embargo, para grandes cantidades de datos se hace una aproximación.

- activation="relu"
- solver='lbfgs'
- alpha=0.1
- random_state=10
- hidden_layer_sizes=(12)

Neural net con dos capas ocultas de 24x24 neuronas

Los parámetros específicos utilizados fueron los siguientes:

- activation="relu"
- solver='lbfgs'
- alpha=0.01
- random_state=10

- hidden_layer_sizes=(24, 24)

Resultados

Desempeño de los modelos: la siguiente tabla presenta la precisión de cada modelo.

MODEL	**ACCURACY**
Modelo base (aleatorio)	**24.72%**
Naive Bayes	**90.63%**
Decision tree	**98.68%**
Random Forest	**97.63%**
Linear SVM	**88.54%**
Neural net una capa oculta	**76.81%**
Neural net dos capas ocultas	**80.40%**

AUC por modelo:

MODEL	**AUC**
Modelo base (aleatorio)	**50%**
Naive Bayes	**91%**
Decision tree	**99%**
Random Forest	**98%**
Linear SVM	**90%**
Neural net una capa oculta	**82%**
Neural net dos capas ocultas	**84%**

Matriz de confusión:

MODEL	**Confusion matrix: [[tn, fp] [fn, tp]]**
Naive Bayes	**[[1674 185] [84 929]]**
Decision tree	**[[1837 22] [16 997]]**
Random Forest	**[[1807 52] [16 997]]**
Linear SVM	**[[1585 274] [55 958]]**
Neural net una capa oculta	**[[1206 653] [13 1000]]**
Neural net dos capas ocultas	**[[1317 542] [21 992]]**

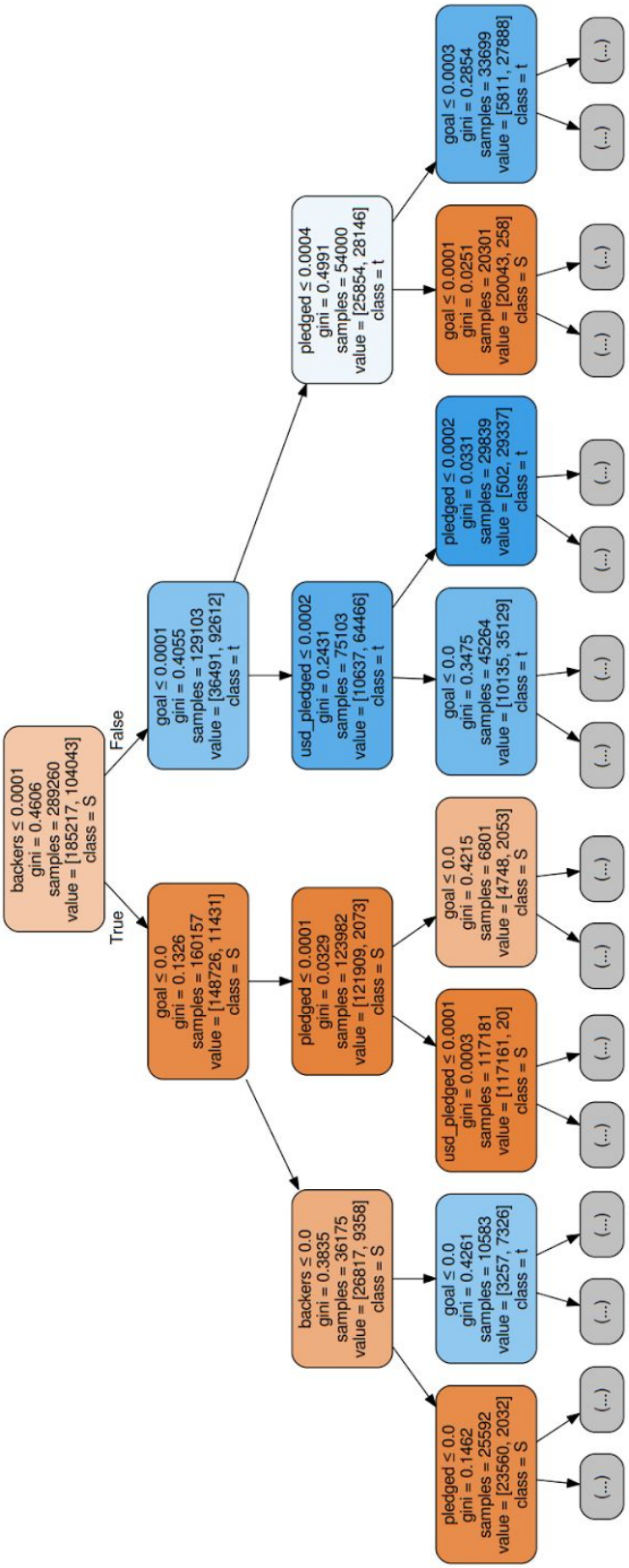
Conclusiones:

Todos los modelos presentaron predicciones con una predicción mayor al 75%. Afortunadamente el mejor modelo fue un árbol de decisión. Si bien al momento de predecir con precisión no es tan importante elegir entre un modelo de redes neuronales y un árbol, para nuestros fines si es importante ya que permite ver qué decisiones toma el algoritmo y qué variables influyen en la determinación del éxito del proyecto (ver nota 1).

En la introducción se mencionó que el objetivo más allá de una buena predicción era dar una herramienta para que los emprendedores del futuro. Con el crecimiento de desarrollos que se están llevando a cabo en este siglo, cualquier persona con una buena idea puede iniciar su propio negocio. Pero la idea no lo es todo, una parte es la idea, otra la implementación y otra diferente levantar capital que te ayude a hacer realidad esas ideas que siguen en papel.

El paso siguiente debería ser analizar, en un equipo multidisciplinario, por qué los factores que determinan el éxito del proyecto lo hacen. Luego crear un simulador que les permita a los futuros emprendedores o desarrolladores de proyectos analizar qué características pueden mejorar en sus propuestas para disminuir la incertidumbre del éxito.

Nota 1: en la siguiente imagen se demuestra la ventaja de los árboles de decisión. Nos permite ver el flujo por el que se determina si un proyecto tendrá éxito o no. En nuestro modelo, el árbol cuenta con 31 niveles de profundidad por lo que sería impráctico mostrarlo aquí, sin embargo, estos son los primeros tres niveles.



Bibliografía:

Kickstarter.com, “About us” https://www.kickstarter.com/about?ref=global-footer Revisado el 15/12/2017

Wikipedia.org, “Kickstarter” https://es.wikipedia.org/wiki/Kickstarter Revisado el 15/12/2017

Techcrunch.com, “Best Overall Startup” https://techcrunch.com/2014/02/10/kickstarter-wins-the-2013-crunchie-for-best-overall-startup-by-leveling-the-maker-playing-field/ Revisado el 15/12/2017

Documentación de SkLearn http://scikit-learn.org/stable/documentation.html Revisado el 15/12/2017

Scikit-learn.org, “”Model evaluation: quantifying the quality of predictions” http://scikit-learn.org/stable/modules/model_evaluation.html Revisado el 15/12/2017

Wikipedia.org, “Hinge loss” https://en.wikipedia.org/wiki/Hinge_loss Revisado el 15/12/2017

Wikipedia.org, “L-BFGS” https://es.wikipedia.org/wiki/L-BFGS Revisado el 15/12/2017

Fuente de los datos del proyecto:

Kaggle, “Kickstarter projects: More than 300,000 kickstarter projects” del usuario Kemical https://www.kaggle.com/kemical/kickstarter-projects/data