Un letrero de color negro

Descripción generada automáticamente con confianza baja

***MODELO DE EVALUACIÓN DE RIESGOS EN STARTUPS***

*INTEGRANTES*

*Julieta Melina Boffelli*

*Juan Martín Marcenaro*

*Fecha de entrega: 7/12*

***Contenidos***

-Descripción del caso de negocio *página 3*

-Descripción de los datos *página 4*

-Descripción de la metodología y

de los objetivos *página 5*

-Consideraciones finales *página 8*

***Descripción del caso de negocio***

***Imagen que contiene Texto

Descripción generada automáticamente***

Una startup es una empresa pequeña o mediana creada hace poco tiempo que surge a partir de un proyecto de negocio que podría ser muy rentable, pero es económicamente arriesgado. Son escalables: es decir, su fin es ser adquiridas o funcionar a partir de financiación externa, empezando con muy pocos recursos. En otras palabras, al contrario de las empresas convencionales, las startups entran rápidamente al mercado buscando nuevas inversiones de capital, por ello sus costes de funcionamiento tienden a ser pequeños, así como su personal suele ser reducido, a menudo contando solo con quienes tuvieron la idea del proyecto.

Si bien hay muchos casos de startups que han acabado siendo muy redituables (como Uber o Spotify) no hay que ignorar que implican un riesgo muy alto. Se puede tener una idea muy innovadora, pero ¿Cómo sabemos que es realmente rentable? ¿Cómo pueden saber los inversores a qué emprendimientos es conveniente darles capital para no perderlo? En base a estas preguntas realizamos nuestro análisis.

Para ello, nos basamos en los datos del Kaggle titulado “Startup Success Prediction: Can you predict if a start-up will succeed or fail?”[[1]](#footnote-1) y a lo largo del presente trabajo describiremos nuestros objetivos y metodología.

***Descripción de los datos***

Nuestro dataset cuenta originalmente con 49 columnas y 923 filas de distintas startups estadounidenses entre 1985 y 2012 que han sido cerradas o adquiridas. A continuación, se detallan las columnas utilizadas.

|  |  |
| --- | --- |
| Columna | Descripción |
| state\_code | Contiene la abreviación del estado al que pertenecen. |
| city | Contiene la ciudad en la que fueron fundadas. |
| name | Contiene el nombre de la startup. |
| status | Contiene las categorías “acquired” en aquellos casos donde la empresa fue absorbida por otra, en cuyo caso es exitosa, y “closed” en los casos en los que la startup no triunfa y cierra. |
| labels | Contiene un 1 en aquellas filas donde la startup tuvo éxito y un 0 en las que han cerrado. |
| founded\_at | Fecha en la que comienza la empresa en formato M/D/A. |
| closed\_at | Si la startup cerró, muestra la fecha en la que cierra la startup en el mismo formato. De otra forma, contiene NaN. |
| first\_funding\_at, last\_funding\_at | Estas columnas muestran, en el mismo formato de fecha, cuándo la startup recibió su primera y su última inversión respectivamente. Si no corresponde, muestra NaN. |
| age\_first\_funding\_year, age\_last\_funding\_year | Muestran cuántos años llevaba la empresa al momento de su primera y su última inversión respectivamente. La “edad” se muestra en números decimales. |
| age\_first\_milestone\_year, age\_last\_milestone\_year | Contienen la cantidad de años que llevaba la empresa al alcanzar su primera meta de inversiones. La edad tiene el mismo formato. |
| relationships | Muestra la cantidad de corporaciones con las que tiene contacto. |
| funding\_rounds | Contiene la cantidad de rondas de financiamiento por las que ha pasado la startup. |
| funding\_total\_usd | Contiene la inversión total en dólares en ese momento (sin contar devaluaciones). |
| milestones | Cantidad de hitos alcanzados, es decir, cuantas metas de inversión alcanzó. |
| is\_CA, is\_NY, is\_MA, is\_TX, is\_otherstate | Estas columnas indican, mediante un 1 para indicar que sí y un 0 para indicar que no, si una startup provenía del estado respectivo. |
| is\_software, is\_web, is\_mobile, is\_enterprise, is\_advertising, is\_gamesvideo, is\_ecommerce, is\_biotech, is\_consulting, is\_othercategory | Estas columnas muestran, con la misma estructura de las anteriores, de qué era la startup. |
| has\_VC | Contiene un 0 en aquellos casos que no han recibido Venture Capital, lo que es un tipo de inversión de riesgo. |
| has\_angel | Un ángel es un individuo que decide invertir una cantidad importante de capital en la startup; a diferencia de la categoría anterior, es una sola persona y no una entidad. La columna muestra un 1 en aquellos casos que sí tuvieron ángel y 0 en los que no. |
| has\_roundA, has\_roundB, has\_roundC, has\_roundD | Las rondas de inversión en las startups son escalares, por lo cual, típicamente se busca una inversión más grande pensando en una meta más ambiciosa mientras más se avanza. Estas columnas dividen los casos en 1 y 0. |
| avg\_participants | Cantidad promedio de participantes. |
| is\_top500 | ¿Ha estado la empresa en el top 500 alguna vez? Si la respuesta es sí, tiene un 1. De lo contrario, tiene un 0. |

Se encontraron algunas columnas con información repetida, en algunos casos expresada de distinta forma y en otros eran columnas exactamente iguales. Lo segundo sucedió con las columnas state\_code y state\_code.1, por lo que nos quedamos con una de ellas tras comprobar que contenían los mismos datos. Lo primero sucedió con las columnas labels y status: contenían los mismos datos, pero la primera los mostraba en formato de 1 y 0 y la segunda asignaba las categorías de acquired y closed. No eliminamos ninguna de las columnas, pero trabajamos más con Status en la fase de EDA y en la aplicación del modelo utilizamos labels por motivos de conveniencia.

Las columnas que sí borramos definitivamente que estaban presentes en el dataset original fueron:

**-***Unnamed: 0* por tener datos que no parecen tener sentido ni utilidad.

**-***zip\_code* y *Unnamed: 6* por tener información sobre la locación de la empresa que es poco relevante para el análisis.

**-***object\_id* e *id* porque son columnas duplicadas con información no útil.

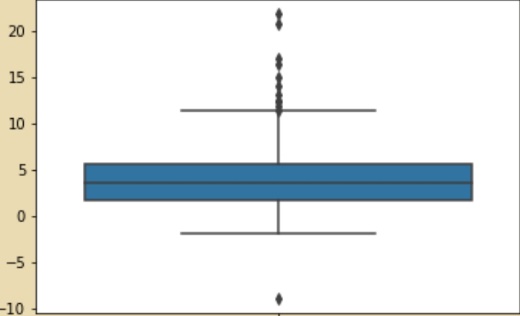
Otra modificación en el dataset original que hicimos fue modificar la columna de founded\_at para que sea de un tipo más trabajable; también, como teníamos información de décadas distintas, dejamos solo el año ya que el día y el mes no nos pareció muy influyente en nuestro análisis.

***Descripción de la metodología y los objetivos***

* ***Primera fase: EDA (Exploratory Data Analysis)***

Al inicio del proyecto, lo primero que nos propusimos fue intentar comprender el estado de los datos: aquí entran el equilibrio de categorías, cantidad de nulos o datos sin sentido, varianza, etc. Una vez que eso estuvo cubierto, la siguiente meta fue intentar encontrar algunas particularidades de nuestra temática elegida.

El primer paso fue calcular la incidencia de los datos nulos en nuestro dataset. De mayor a menor, las columnas con datos nulos fueron closed\_at, Unnamed: 6, age\_first\_milestone, age\_last\_milestone y state\_code.1. Como se ha dicho anteriormente, tanto Unnamed: 6 como state\_code.1 serían borradas luego, y las columnas restantes tenían una explicación sencilla para la presencia de datos nulos: si la empresa nunca había alcanzado cierta meta o no había cerrado, no había razón para que contengan datos. Así que no nos encontramos con grandes adversidades en esta cuestión.

Por otro lado, cuando comenzamos a llevar a cabo el análisis univariado descubrimos, tras realizar boxplots, que muchas las columnas relacionadas con la edad de la empresa cuando algo sucedía algo extraño: tenían datos negativos. Esto era claramente un dato erróneo, así que estas filas fueron desestimadas en los análisis y aplicaciones en las que eran necesarias.

*Boxplot generado a partir de la columna age\_last\_funding\_year*

En el análisis bivariado se dividieron las startups según el año en el que fueron fundadas y se realizó un gráfico de barras que contaba la cantidad que habían acabado cerrando y siendo adquiridas, respectivamente. Descubrimos así que en el período 2005-2009 este tipo de empresas tuvieron una marcada explosión, ya que muchísimas fueron adquiridas, y, en los datos de años anteriores, los números tendían a ser mucho más desalentadores. Como tiene sentido, después del boom del período 2005-2009, comenzaron a acrecentarse las startups que cerraban.

Se realizó también un PCA (Principal Component Analysis o Análisis de Componentes Principales) en las columnas age\_first\_funding\_year, age\_last\_funding\_year, age\_first\_milestone\_year, age\_last\_milestone\_year, relationships, funding rounds, funding\_total\_usd, milestones y avg\_participants, ya que eran las únicas con datos numéricos en el dataset usado. Como se ha dicho, se han excluido de este análisis las filas con datos sin sentido en las columnas de edades, así como los datos nulos; a las filas restantes se las normalizó. Como resultado, se descubrió que 5 componentes explicaban casi el 90% de la varianza.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* ***Segunda fase: búsqueda y aplicación del modelo***

Por los requerimientos del proyecto y las características de nuestros datos, debíamos elegir un algoritmo de clasificación de aprendizaje supervisado. Nuestra variable target fue, evidentemente, la columna labels, ya que nuestra idea es predecir qué startups acaban teniendo éxito y cuáles son inversiones de riesgo que no servirán. Por la misma naturaleza de nuestro objetivo, tomamos dos métricas: el porcentaje de aciertos (Accuracy Score) y los resultados de la matriz de confusión, tomando como mejores aquellos resultados que tenían menos falsos positivos (empresas diagnosticadas como exitosas que en realidad fracasaron). Teniendo esto en cuenta, pasaremos a describir brevemente las opciones que probamos y los resultados que obtuvimos en cada una.

*Árbol de decisión*

Tuvo un 84% de aciertos sobre el set de entrenamiento y casi un 78% en el de evaluación. Al aplicar la matriz de confusión, se puede observar que no es tan apropiado: tuvo 41 falsos positivos, junto con 158 verdaderos positivos, 38 verdaderos negativos y 15 falsos negativos.

*Random Forest*

Si evaluamos su desempeño mediante los aciertos, este algoritmo presentó un desempeño de 86% de aciertos en el set de entrenamiento y 81% en el set de evaluación tras aplicar Grid Search (el mejor hasta ahora). En lo que sí mejoró en comparación al anterior algoritmo, aunque levemente, fue en su desempeño en la matriz de confusión: tuvo 40 falsos positivos, 166 verdaderos positivos, 7 falsos negativos y 39 verdaderos negativos.

*XG Boost*

Se probó el algoritmo XGBoost. Su desempeño en el Accuracy Score fue de casi el 87% en el set de entrenamiento, aunque descendió a un 80% en el set de evaluación. En cuanto a la matriz de confusión, devolvió 36 falsos positivos (la menor cantidad hasta ahora), 159 verdaderos positivos, 14 falsos negativos y 43 verdaderos negativos.

*Regresión logística*

Por último, la regresión logística es un análisis de regresión que predice el valor de una variable categórica a partir de las variables independientes. Pensábamos que iba a ir muy bien con el problema a tratar porque es muy usada para predecir eventos ocurridos en función de ciertos factores, sin embargo, no tuvo resultados sobresalientes, aun habiendo aplicado el método de Grid Search en todos los casos. Su accuracy score en entrenamiento y evaluación fue del 78% y 79% respectivamante. En cuanto a la matriz de confusión se obtuvieron 38 falsos positivos, 16 falsos negativos, 157 verdadero positivos y 41 verdadero negativos.

*Modelo elegido*

Creemos que, si bien el accuracy score más elevado lo obtuvo el modelo de Random Forest, el modelo que mejor se ajusta a los datos es el de XG-Boost por ser el menos sensible a los falsos positivos y al mismo tiempo tener un accuracy score relativamente bueno. Para este problema en particular, es necesario un modelo que mejor identifique las startups que podrían llevar al fracaso ya que se trata de negocios riesgosos y de mucha inversión para aquellos que aportan el capital. El modelo generado servirá como una herramienta de decisión, que a través de las características de una dada startup permita predecir su éxito o fracaso.

***Consideraciones finales***

Creemos que nuestro trabajo ha cumplido de manera modesta con el objetivo que nos habíamos planteado: predecir qué startups del dataset fueron exitosas y cuales no. Sin embargo, a partir del EDA descubrimos que el futuro de una startup no está condicionado sólo por la inversión que recibe o de qué lugar proviene: el estado del mercado genera fluctuaciones a las que estos emprendimientos de alto riesgo económico son, claramente, muy sensibles. De todas maneras, el modelo actual, aunque limitado, puede ser usado por inversores para orientarse ante qué startups apadrinar.

Aun así, tenemos la certeza de que, teniendo una mayor cantidad de datos para un período más reducido se podrían llegar a conclusiones más útiles. En adición, si los resultados de ese análisis en distintos períodos se complementa con un estudio del estado del mercado en las respectivas divisiones temporales, se podría a arribar a conclusiones muchísimo más sólidas. Aún así, nuestro proyecto tuvo la intención de ser un primer paso dentro de esta cuestión, no de solucionar algo tan complejo: es por eso que creemos, aún más después de haber finalizado el proyecto, que el estudio de esta temática tiene mucho potencial.

1. Disponible en https://www.kaggle.com/manishkc06/startup-success-prediction [↑](#footnote-ref-1)