Máster Universitario en Investigación en Inteligencia Artificial

Ciencia de Datos y Aprendizaje Automático

Proyecto `Kaggle`

Retención de clientes

Manuel Germán Morales y Juan Carlos Alfaro Jiménez

{mgerman@ujaen.es, juancarlos.alfaro@uclm.es}

Noviembre 2025

Índice

1. Kaggle

Kaggle (www.kaggle.com) es una plataforma donde se reúnen miles de personas con interés o experiencia en el análisis de datos, ofreciendo la posibilidad de competir para resolver requisitos estratégicos que presentan los grandes datos de las empresas a cambio de dinero. Empresas y compañías de todo el mundo exponen sus problemas y sus retos en esta plataforma y la comunidad de científicos de datos compite para crear las mejores soluciones y los mejores modelados teóricos.

En la plataforma se presenta cualquier tipo de problema que pueda encontrarse en los distintos campos del mundo real, tales como servicios financieros, energía, tecnologías de la información, etc.

El enfoque *crowdsourcing* utilizado se debe a la existencia de una cantidad indefinida de posibles soluciones y estrategias que se pueden aplicar a un problema complejo de modelado predictivo donde no es posible saber con antelación la técnica o la estrategia que será más adecuada y más eficaz.

Fundada por el economista australiano Anthony Goldbloom, la inspiración para crear Kaggle proviene en parte de una competición creada por Netflix entre 2006 y 2009. La empresa de alquiler de películas ofrecía un millón de dólares al equipo que fuera capaz de mejorar la precisión de su sistema de recomendación de títulos en un 10 %.

1.1. Registro

El primer paso para el proyecto es registrarse en la plataforma para crear una nueva cuenta de usuario. Una vez se ha realizado el registro, entra en la plataforma y accede al apartado "*Competitions*" donde encontrarás una lista con todas las competiciones, tanto activas como completadas. Busca la competición titulada como "*Retención de clientes para una entidad financiera*" (Figura 1) o accede directamente a través del siguiente enlace de invitación: https://www.kaggle.com/t/72ef18031972498a9b9d0917b51c7564.

Figura 1: Cabecera de la competición *Retención de clientes para una entidad financiera*

Para participar en la competición y descargar los datos del problema deberás aceptar las reglas de la misma.

2. Descripción del problema

La fuga de clientes constituye uno de los mayores desafíos para las instituciones financieras, ya que adquirir un nuevo cliente suele resultar mucho más costoso que retener a uno existente. Además, cuando un cliente decide cerrar su cuenta, la entidad pierde de manera irreversible la oportunidad de desplegar campañas de fidelización o de ofrecer productos alternativos que podrían haber evitado su salida. Este fenómeno no solo afecta los ingresos directos del banco, sino también su estabilidad a largo plazo, al perder la confianza y la base de clientes. En este contexto, la capacidad de predecir con antelación qué clientes tienen mayor probabilidad de abandonar la entidad se convierte en una herramienta estratégica esencial permitiendo focalizar esfuerzos, optimizar recursos en campañas de retención y, en definitiva, mejorar la relación con los clientes y la sostenibilidad del negocio.

Con respecto al conjunto de datos, este está formado por 10000 instancias y 14 variables (incluyendo la variable clase a predecir), las cuáles se centran en la descripción de multitud de aspectos de cada uno de los clientes de un banco. Estas variables recogen información detallada de ellos, tanto demográfica como financiera, tales como edad, país de residencia, género, saldo, antigüedad en la entidad, número de productos contratados, posesión de tarjeta de crédito, condición de miembro activo, nivel de ingresos estimado, entre otras. La variable objetivo (*Exited*) indica si el cliente abandonó (1) o permaneció en el banco (0).

La tarea del alumno es predecir si los clientes cerraron o no sus cuentas. Para comenzar a trabajar, en la pestaña "*Data*" encontraremos los siguientes ficheros: `train.csv`, `test.csv` y `sample_submission.csv`. El conjunto de datos de entrenamiento se encuentra en `train.csv` y consta de 8000 instancias. Por su parte, el conjunto de datos de prueba es `test.csv`, está formado por 2000 instancias y se utilizará para realizar la predicción. Por último, el fichero `sample_submission.csv` proporciona un formato de muestra para el envío de resultados a la plataforma `Kaggle`. Recuerda que el objetivo de este proyecto es demostrar las habilidades aprendidas por el estudiante durante el curso, por lo que **debes** ir más allá que la mera aplicación de un algoritmo sobre el conjunto de datos dado.

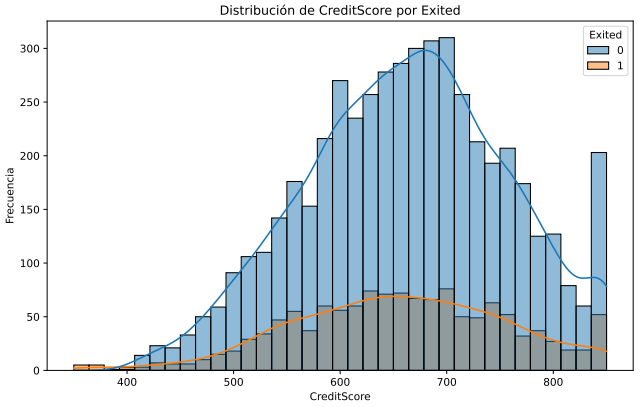
3. Exploración y preprocesamiento de datos

Una vez descargados los datos, es necesario realizar un **extenso análisis descriptivo del conjunto de datos de entrenamiento**, así como un proceso de inspección, limpieza y transformación de datos con el objetivo de resaltar información útil para la fase de modelado. Este análisis nos permitirá controlar la presencia de valores faltantes o la presencia de posibles errores en la fase de introducción de los datos. También nos proporcionará una idea inicial de la forma que tienen los datos (distribución, parámetros de dispersión, etc.), así como las relaciones entre los distintos atributos.

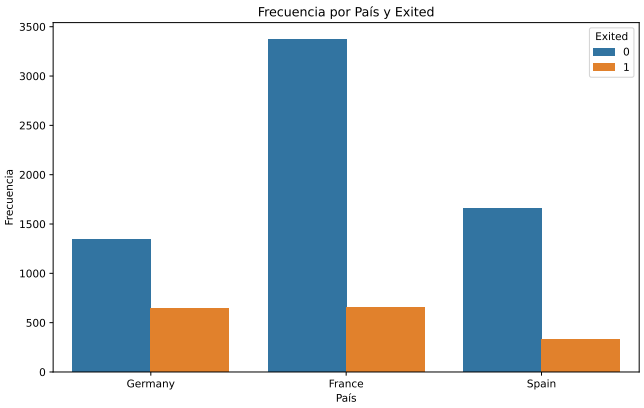
La función principal de esta exploración de los datos es que se utilicen las distintas gráficas, coeficientes, y estadísticas utilizadas como base de cualquier decisión tomada para la generación de los consiguientes modelos de clasificación. Como posibles puntos de partida, se puede analizar los clientes que cerraron su cuenta en base a su calificación creditícea (*CreditScore*) o el país de residencia (*Geography*), como se muestra en la Figura 2.

3.1. Consejos

- Recuerda que estamos trabajando sobre el conjunto de datos de entrenamiento y que todas aquellas operaciones de limpieza, transformación, creación de nuevos atributos, etc. que realicemos debemos también aplicarlas sobre el conjunto de datos de prueba. **Para evitar fugas de datos, se debe usar siempre la información del conjunto de datos de entrenamiento. Para ello, se recomienda usar el concepto de *pipeline***. En los siguientes enlaces se proporciona más información para las principales librerías de aprendizaje automático:
  - `Scikit-learn`: https://scikit-learn.org/stable/modules/compose.html
  - `Caret`: https://recipes.tidymodels.org/articles/recipes.html
  - `Weka`: https://waikato.github.io/weka-wiki/
- Analiza la existencia de posibles valores anómalos, extremos o inusuales.
- Comprueba si en el conjunto de datos existen valores faltantes. Muchas técnicas **no** pueden procesar observaciones con valores faltantes. Decide qué hacer con los valores faltantes en los atributos (eliminación,



(a) `CreditScore`



(b) `Geography`

Figura 2: Distribución de los diferentes atributos condicionados a los valores de la variable clase.

transformación, imputación, etc.).

- Explora correlaciones entre atributos numéricos y la variable a predecir.
- Se pueden transformar las variables categóricas en numéricas (p.e., codificación *one-hot*), o numéricas en categóricas (p.e., discretización), aplicando distintas aproximaciones.
- Para obtener un mejor ajuste de los datos, es necesario realizar un análisis más detallado de cada variable categórica.
- Puede ser de utilidad la creación de nuevos atributos a partir de atributos ya existentes que permitan mejorar la descripción de los datos, así como reducir su dimensionalidad.
- También puede ser interesante crear nuevos atributos basados en la interacción de variables altamente correlacionadas.

4. Predicción y evaluación

Una vez hemos realizado nuestro análisis sobre los datos, incluyendo la creación del *pipeline* de preprocesamiento con la limpieza, transformación y generación de nuevas variables interesantes para nuestro estudio, pasamos a

CustomerId	Exited
11111	1
22222	0
33333	1
44444	0
55555	0

Cuadro 1: Muestra del fichero `.csv` para su evaluación en `Kaggle`

la fase del modelado. **De nuevo, se recomienda que el modelo se introduzca, junto con el preprocesamiento, en un *pipeline* de modelado con el objetivo de garantizar que el entrenamiento e inferencia se realiza de una manera metodológica correcta**.

Una vez se ha entrenado el modelo, se puede obtener una evaluación de la predicción en `Kaggle`. Para ello, es necesario subir los resultados obtenidos utilizando el conjunto de datos de prueba a la misma. De esta manera, utilizamos el modelo para realizar la predicción y guardamos el resultado en un fichero `.csv` con el formato que se muestra en la Tabla 1.

Para poder evaluar tu predicción debes volver a la competición en `Kaggle` y subir el archivo `.csv` con la predicción y enviar. Tras unos segundos, y si no hay errores, se obtendrá la **puntuación F1** o *F1-score* (métrica de rendimiento utilizada en `Kaggle` como evaluación). Para mejorar los resultados, debemos trabajar más tanto en la fase de preprocesado como en la de modelado.

Se debe tener en cuenta que sólo se puede subir un máximo de **5 ficheros de predicciones al día**.

En el fichero `ejemplo_lda.html` adjunto se proporciona un ejemplo ilustrativo de todo este proceso usando un `pipeline` de preprocesamiento y una técnica de modelado sencilla como es un análisis discriminante lineal. Para ello, se ha utilizado la librería `scikit-learn` en `Python`.

5. Objetivo

El objetivo de esta práctica es muy sencillo de explicar: **Construir un modelo de aprendizaje automático que obtenga una estimación lo más precisa posible de la variable `Exited` a partir del resto.**

6. Criterios de evaluación

El proyecto debe realizarse en el lenguaje de programación de preferencia. Como entregable de este proyecto se realizará una memoria de **máximo 10 páginas** describiendo el proceso realizado en cada una de las fases de proyecto. Puede servir como guía (no exhaustiva):

- Introducción y descripción del problema
- Comprensión y preprocesado de datos:
  - Exploración y visualización de datos
  - Limpieza (valores faltantes, extremos, etc.)
  - Procesado de variables numéricas (estudios de correlaciones, discretizaciones u otras transformaciones, etc.)

  - Procesado de variables categóricas (asociaciones y dependencias, numerizaciones, etc.)
  - Construcción, formateo y estandarización de variables (creación de nuevas variables, normalizaciones, etc.)
  - Selección horizontal o vertical (selecciones de atributos basadas en filtros, modelos, correlaciones, pesos, etc.)
- Modelado (técnicas de aprendizaje utilizadas, construcción, comparativas, hiperparámetros, etc.)
- Evaluación (criterio de evaluación, resultados, comparativas, etc.)

Finalmente, se deberá incluir también:

- Conclusiones (análisis de resultados, lecciones aprendidas tras la realización del proyecto, etc.)
- Resultado en `Kaggle` (usuario y valor de la métrica de rendimieto obtenida mediante una captura de pantalla)

Se valorará muy positivamente la claridad de la memoria y la capacidad de síntesis. Aunque se hayan probado muchas cosas, los alumnos deben decidirse por un modelo únicamente, que es el que se evaluará en la plataforma `Kaggle` (generalmente el que hayan conseguido colocar más arriba, pero podría ser otro de su elección). **Es fundamental razonar y justificar todas las decisiones tomadas por el alumno en cada una de las fases**. Se deberá especificar claramente el usuario utilizado y el resultado obtenido para su comprobación mediante una captura de pantalla.

Por otra parte, además de la memoria, **se evaluará al alumno por medio de una entrevista por videoconferencia acerca del desarrollo del proyecto**. En la misma, se espera que se describirá brevemente el desarrollo realizado y deberá responder acerca de cualquier particularidad del proyecto que se considere relevante. Una vez finalizada la fecha de entrega de la memoria, el profesorado encargado de la evaluación del proyecto final se pondrá en contacto para concertar una cita.

Se deberán respetar todas las normas de `Kaggle` y no usar otros usuarios adicionales al de la práctica para conseguir sobrepasar el límite de subidas por día que permite la plataforma.

La comunidad de usuarios de `Kaggle` proporciona una gran ayuda para la resolución de los distintos proyectos por medio de sus foros y libretas. Se recomienda la utilización de los mismos tanto para comenzar con el proyecto como para resolver dudas o coger ideas. **Cualquier pieza de código utilizada que provenga de cualquiera de estas fuentes debe ser citada en la evaluación**.

☞ **Adjunta tu código mediante un enlace a plataformas que te permitan alojarlo como *GitHub*, *GitLab* o similares, pero NUNCA lo copies directamente en tu memoria.**

Para la nota final se valorará que:

- La predicción del modelo haya sido evaluada en `Kaggle`. Cuanto mayor sea la tasa de acierto, mejor será la valoración por este concepto.
- Se haya trabajado en el preprocesado (análisis, limpieza y transformación) de los datos y su comprensión.
- Se haya trabajado en el modelado intentando mejorar el resultado.
- Se haya realizado una validación correcta.

- Se responda adecuadamente a las cuestiones planteadas y que de las respuestas no haya dudas de que se ha realizado la práctica y que se sabe manejar mínimamente las herramientas que se hayan utilizado.