Práctica 2 (35% nota final)

Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas. Para hacer esta práctica tendréis que trabajar en grupos de 2 personas. Tendréis que entregar un solo archivo con el enlace Github (https://github.com) donde se encuentren las soluciones incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github. Aunque no se trata del mismo enunciado, los siguientes ejemplos de ediciones anteriores os pueden servir como guía:

- Ejemplo: https://github.com/Bengis/nba-gap-cleaning
- Ejemplo complejo (archivo adjunto).

Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el

problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

Descripción de la Práctica a realizar

Este proyecto ha sido desarrollado y elaborado por:

- Mónica Alexandra Gómez Martínez
- David Francisco Dávila Ortega

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com). Algunos ejemplos de dataset con los que podéis trabajar son:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset elegido se llama 'Credit Card customers' y puede descargarse desde la plataforma Kaggle. Su importancia radica en que trata de resolver un frecuente problema de negocios. Este conjunto de datos contiene la información de una agencia bancaria, donde el administrador desea conocer el o los motivos por los cuales están perdiendo clientes, así como prevenir el así llamado 'Credit Card Churning'.

Habiendo identificado los clientes que potencialmente desean prescindir de algún servicio bancario, la agencia podría anticipar estrategias para lograr que dichos clientes se queden por más tiempo.

Este dataset contiene información sobre 10,127 clientes. Entre los datos descriptores se encuentran 23 variables que son:

0 CLIENTNUM: de tipo entero y sin valores nulos
1 Attrition_Flag: de tipo entero y sin valores nulos
2 Customer_Age: de tipo entero y sin valores nulos
3 Gender: de tipo objeto y sin valores nulos
4 Dependent_count: de tipo entero y sin valores nulos
5 Education_Level: de tipo objeto y sin valores nulos
6 Marital_Status: de tipo objeto y sin valores nulos
7 Income_Category: de tipo objeto y sin valores nulos
8 Card_Category: de tipo objeto y sin valores nulos
9 Months_on_book: de tipo entero y sin valores nulos
10 Total_Relationship_Count: de tipo entero y sin valores nulos
11 Months_Inactive_12_mon: de tipo entero y sin valores nulos
12 Contacts_Count_12_mon: de tipo entero y sin valores nulos
13 Credit_Limit: de tipo flotante y sin valores nulos
14 Total_Revolving_Bal: de tipo entero y sin valores nulos
15 Avg_Open_To_Buy: de tipo flotante y sin valores nulos
16 Total_Amt_Chng_Q4_Q1: de tipo flotante y sin valores nulos
17 Total_Trans_Amt: de tipo entero y sin valores nulos
18 Total_Trans_Ct: de tipo entero y sin valores nulos
19 Total_Ct_Chng_Q4_Q1: de tipo flotante y sin valores nulos
20 Avg_Utilization_Ratio: de tipo flotante y sin valores nulos
21 Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1: de tipo objeto y sin valores nulos
22 Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2: de tipo flotante y sin valores nulos

Este dataset se encuentra completamente en el idioma inglés y se trabajará como tal sin hacer traducciones de variables o información. Las modificaciones realizadas

(preprocesamiento) se encuentran comentadas dentro del archivo Rmarkdown (PRA2_Limpieza.Rmd).

2. Integración y selección de los datos de interés a analizar.

Para ello se trabajó con Rmarkdown en Rstudio:

Selección de los datos de interés.- Para el presente proyecto se utilizarán todas las variables presentes en el juego de datos a excepción de las dos últimas que no serán de utilidad en los análisis que se planean hacer a posteriori.

Análisis gráfico descriptivo.- De manera general se observa que las variables cuantitativas presentan una alta dispersión en sus datos, en cuanto a las variables cualitativas, por ejemplo, se puede observar que el banco en cuestión presenta una mayor cantidad de clientes con cuentas activas que inactivas, existen más mujeres que hombres, la mayoría están solteros o casados con estudios terminados y que usan la tarjeta de crédito categoría “Blue”.

3. Limpieza de los datos.

El proceso paso a paso se encuentra en el archivo: ‘PRA2_Limpieza.html’, los resultados se encuentran en el archivo ‘BankChurners.csv’ y el dataset procesado se denomina ‘‘BankChurners_clean.csv’

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Este dataset en particular presenta una completitud del 100%. En el caso de haber encontrado elementos vacíos o nulos es conveniente conocer la fuente de los datos antes de proceder a eliminarlos ya que se puede estar perdiendo valiosa información. Lo que se hubiera hecho en este caso es una imputación de valores vacíos/nulos y, en medida de lo posible, evitar la eliminación de algún valor.

3.2. Identificación y tratamiento de valores extremos.

La distribución no es simétrica para la mayoría de variables, y es simétrica en los casos de: "Customer_Age", "Dependent_count" y "Months_on_book".

Para la mayoría de variables se ha optado por mantener los valores originales con excepción de aquellos outliers encontrados para las variables: “”Customer_Age” y Total_Trans_Ct”.

Un análisis variable por variable se detalla en el archivo Rmarkdown (.Rmd).

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En nuestro análisis se plantean tres diferentes grupos de interés: los grupos correspondientes al género, el nivel de educación y estado marital. Esta información se encuentra más detallado dentro del archivo Rmarkdown (.Rmd)

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Es de principal importancia dependiendo del modelo a aplicar, el realizar la correspondiente comprobación de normalidad y homogeneidad de la varianza en nuestro conjunto de datos. Por lo cual, para realizar la comprobación de nuestras variables sobre una población normalmente distribuida, se utilizará la prueba de normalidad de Anderson-Darling.

Teniendo en cuenta en $\alpha=0.05$, si obtenemos un p-valor superior al alfa mencionado anteriormente, entonces no rechazaremos nuestra hipótesis nula y los valores vendrán de una distribución/población normal.
En nuestro caso, las variables que no siguen una distribución normal:

- Attrition_Flag
- Customer_Age
- Dependent_count
- Months_on_book
- Total_Relationship_Count
- Months_Inactive_12_mon
- Contacts_Count_12_mon
- Credit_Limit
- Total_Revolving_Bal
- Avg_Open_To_Buy
- Total_Amt_Chng_Q4_Q1
- Total_Trans_Amt
- Total_Trans_Ct
- Total_Ct_Chng_Q4_Q1Avg_Utilization_Ratio

Posterior a ello, realizamos la comprobación sobre la homogeneidad de las varianzas con el test de Fligner.Killen. Para nuestro caso, estudiaremos la homogeneidad en cuanto a los grupos conformados por las siguientes variables: Género, nivel educativo, estado marital, categoría de ingresos y categoría de la tarjeta. Recordemos que la hipótesis nula consiste en que ambas varianzas son iguales

Teniendo en cuenta los resultados, solamente podemos no rechazar la hipótesis nula para las variables: nivel educativo, estado marital y categoría de la tarjeta. Dado que se obtienen que el valor-p asociado a la prueba es mayor que el alfa establecido.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Dentro de las pruebas estadísticas planteadas, se establecieron correlaciones con las variables continuas dentro del conjunto de datos, en los cuales se obtuvieron resultados de correlaciones significativas sin embargo estas correlaciones se denominan de baja correlación ya que el máximo valor que éstas toman está alrededor de 0.30.

Por otra parte también se realizó un modelo de regresión logística para modelar el nivel de deserción de los usuarios ante una empresa, esta información de los modelos y de las correlaciones se encuentra más detallada dentro del Rmarkdown (.Rmd)

5. Representación de los resultados a partir de tablas y gráficas.

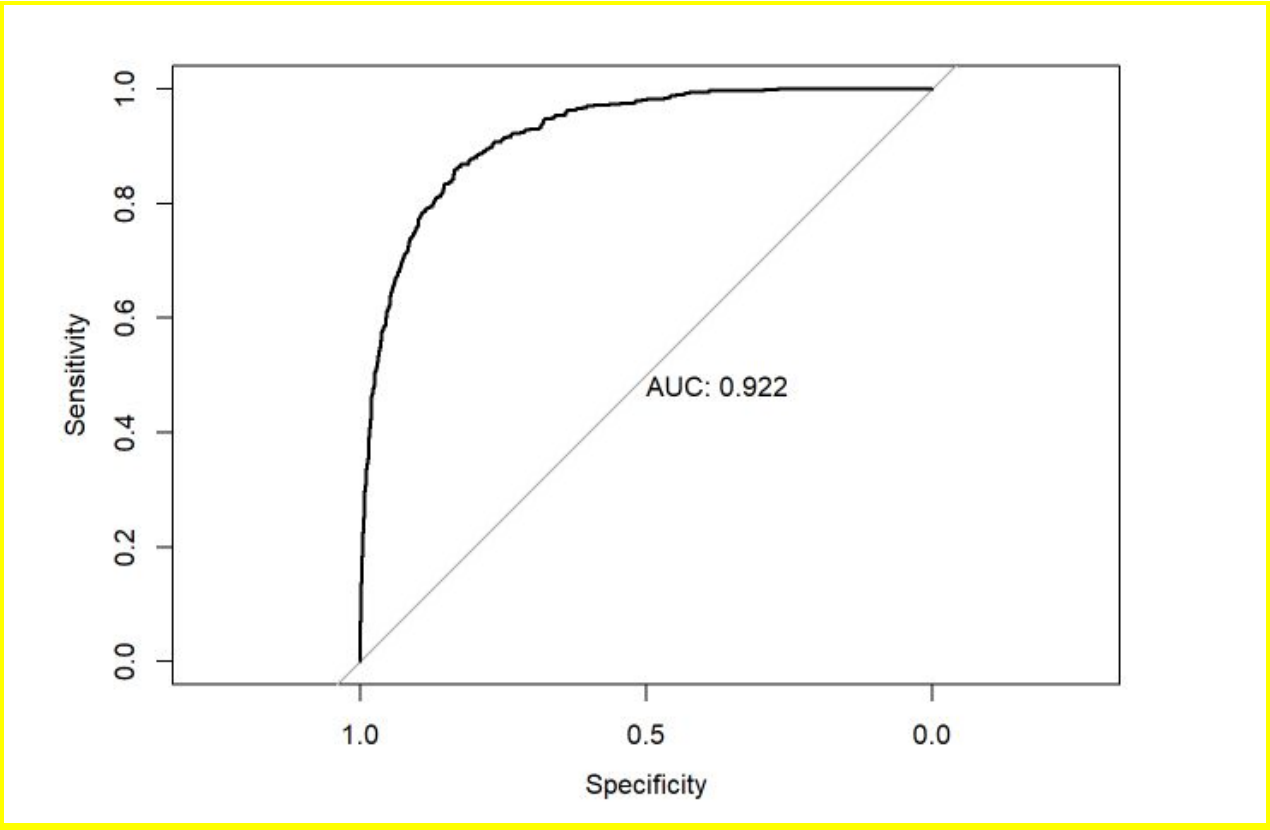
Interpreation de coeficientes (odds)

Una breve interpretación de los resultados obtenidos con los coeficientes u “odds”, puede ser la siguiente:

- Los usuarios que se encuentran solteros aumentan la tasa de abandono en un 11%.
- Dentro de los usuarios que tienen diferentes categoráis de tarjetas se tiene que aquellos que tienen la categoría platino aumentan la tasa de abandono en un 119%.
- Una relación inversa que observamos la tasa de abandono es un un 59% menor si el usario es de género masculino comparado con las usuarias femeninas.

(Intercept)	Customer_Age
89.15	0.99
GenderM	Dependent_count
0.41	1.14
Marital_StatusMarried	Marital_StatusSingle
0.60	1.11
Marital_StatusUnknown	Income_Category$40K - $60K
1.01	0.52
Income_Category$60K - $80K	Income_Category$80K - $120K
0.59	0.74
Income_CategoryLess than $40K	Income_CategoryUnknown
0.59	0.51
Card_CategoryGold	Card_CategoryPlatinum
2.18	2.19
Card_CategorySilver	Total_Relationship_Count
1.13	0.63
Months_Inactive_12_mon	Contacts_Count_12_mon
1.64	1.71
Total_Revolving_Bal	Total_Trans_Amt
1.00	1.00
Total_Trans_Ct	
0.89	

Por otra parte se realiza la gráfica ROC en la cual contiene también el valor de AUC (más conocida como el área bajo la curva ROC), hay que recordar que este valor varia entre 0 y 1, donde un modelo cuyas predicciones son un 100% incorrectas tiene un AUC de 0.0, otras donde sus predicciones son de un 100% entonces su valor de AUC asociado es de 1.0. En nuestro caso, obtuvimos un valor de AUC igual a 0.922 lo cual nos indica que las predicciones que se están haciendo se desvían un poco pero están cerca de pertenecer a un modelo “perfecto”.



6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Teniendo en cuenta los resultados por ejemplo de la exactitud y la precisión presentada dentro del código se recomendaría el poder utilizar nuevas variables que nos ayuden a encontrar relaciones escondidas dentro de los datos y generen un mejor modelo con una exactitud y precisión más alta. Sin embargo los resultados actuales que se tienen del modelo de regresión logístico son bastante buenos teniendo en cuenta los datos y la cantidad de datos que se obtuvieron para la realización de la práctica. Otra opción que tendíamos a futuro, sería plantear otros tipos de modelos que se comparen con el modelo actual, tal como lo puede ser un árbol de decisión, clústeres e inclusive redes neuronales para manejar las relaciones complejas que no se pueden detectar utilizando otros modelos anteriormente mencionaldos.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

Project Link:https://github.com/wickedlexie/AnalisisPRA2

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.
- Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media.
- Wes McKinney (2012). *Python for Data Analysis*. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la

siguiente:

- Los apartados 1, 2 y 6 valen 0,5 puntos.
- Los apartados 3, 5 y 7 valen 2 puntos.
- El apartado 4 vale 2,5 puntos.

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

Formato y fecha de entrega

Durante la semana del 21 al 25 de diciembre el grupo podrá entregar al profesor una entrega parcial opcional. Esta entrega parcial es muy recomendable para recibir asesoramiento sobre la práctica y verificar que la dirección tomada es la correcta. Se entregarán comentarios a los estudiantes que hayan efectuado la entrega parcial pero no contará para la nota de la práctica.
En la entrega parcial los estudiantes deberán entregar por correo electrónico, al profesor

encargado del aula, el enlace al repositorio Github con el que hayan avanzado.

En referente a la entrega final, hay que entregar un único fichero que contenga el enlace Github, el cual no se podrá modificar posteriormente a la fecha de entrega, donde haya:

1. Una Wiki con los nombres de los componentes del grupo y una descripción de los ficheros.
2. Un documento PDF con las respuestas a las preguntas y los nombres de los componentes del grupo. Además, al final del documento, deberá aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación de que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	**Firma**
Investigación previa	**DFDO (dfdavila)** **Mónica Alexandra Gómez Martínez**
Redacción de las respuestas	**DFDO (dfdavila)** **Mónica Alexandra Gómez Martínez**
Desarrollo código	**DFDO (dfdavila)** **Mónica Alexandra Gómez Martínez**

3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se debe entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59** del día **5 de enero**. No se aceptarán entregas fuera de plazo.