UNIVERSIDAD NACIONAL DE LA MATANZA

FUNDAMENTOS DE LA CIENCIA DE DATOS E INTELIGENCIA DE NEGOCIOS

Evaluación de Aprendizaje N° 2

Noviembre de 2020

Objetivos

- Desarrollar hipótesis sobre el conjunto de datos utilizando técnicas de análisis exploratorio.
- Comprender los algoritmos utilizados en Machine Learning para problemas de clasificación.
- Desarrollar modelos predictivos para resolver problemas en el ámbito de la ciencia de datos.
- Comunicar los resultados del análisis de datos, a través de técnicas de visualización de información adecuadas e interpretables.

Caso Propuesto

La empresa “Business Prop SRL” contrata nuestros servicios para que le desarrollemos un modelo que permita predecir si las casas vendidas pagan o no comisión, cuando su precio de venta sea superior a un determinado valor.
Para ello, nos comparten un dataset llamado **casas_entrenamiento.csv**, que contiene información de departamentos vendidos en distintos lugares de Argentina y el exterior. Este dataset será el que utilicemos para el entrenamiento del modelo construido.
El dataset de predicción a utilizar es **casas_predecir.csv**, el cual no contiene la etiqueta de la variable clase (por defecto viene indicada como “no paga”). Cada uno de estos datasets pueden descargarlos desde:

`https://raw.githubusercontent.com/unlam-fcdin/UNLaM_FCDIN/master/casas_entrenamiento.csv`
`https://raw.githubusercontent.com/unlam-fcdin/UNLaM_FCDIN/master/casas_predecir.csv`

a. Función de ganancia y criterio de aprobación

Por cada predicción que el modelo acierte, la empresa les pagará $100 en concepto de comisión, pero si fallan y cobran una comisión que no deberían, deben darle a Business Prop SRL una compensación de $50.

La evaluación de aprendizaje se considerará aprobada si la ganancia obtenida por el modelo sobre los datos desconocidos supera los $650.000. Como referencia, la ganancia que calculará el modelo construido es sobre el porcentaje de los datos usados para testing (muestra del 40%), con lo que ese valor puede ser alrededor de $300.000.

Cada 4% más de ganancia que obtenga por encima de la mínima requerida sumará un punto en la nota de la evaluación, si alcanza o supera el 25% la calificación será la máxima.

b. Entregable

Generar y enviar un archivo de salida **casas_entrega.csv** con los casos donde la medición realizada infiera que se pagará comisión. Para ello, deberá utilizar el dataset que se usó para predecir casas_predecir.csv. Enviar también el notebook construido con el trabajo desarrollado.

Se requiere probar al menos dos (2) algoritmos y construir más de un modelo, los cuales deberán estar incluidos en el notebook entregado. Recuerde que el objetivo es lograr mejorar su modelo original.

c. Aclaraciones y sugerencias

Algunas variables pueden generar alta dimensionalidad en el dataset, como ser los campos "title","start_date","end_date","created_on",”l3”,”l4” y ”l5”. Sugerimos comenzar con un modelo que no incluya estos campos o bien analizar qué información interesante de los mismos podría contribuir a la mejora de la predicción. Por otro lado, sugerimos que el campo "id" lo usen como índice, con el comando set index("id").

Fecha de Entrega: 10/11/2020