FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 2: Machine Learning

[75.06 / 95.58] Organización de Datos
Segundo cuatrimestre de 2018

Grupo Datatouille

Alumno	Padrón	Mail
Bojman, Camila	101055	camiboj@gmail.com
del Mazo, Federico	100029	delmazofederico@gmail.com
Hortas, Cecilia	100687	ceci.hortas@gmail.com
Souto, Rodrigo	97649	rnsoutob@gmail.com

https://github.com/FdelMazo/7506-Datos/
https://kaggle.com/datatouille2018/

Curso 01

- Argerich, Luis Argerich
- Golmar, Natalia
- Martinelli, Damina Ariel
- Ramos Mejia, Martín Gabriel

Contents

1 Introducción

Se propone exponer en el siguiente informe el desarrollo del Trabajo Práctico para predecir la probabilidad de que un usuario de Trocafone realice una conversión en un período determinado de tiempo. Con dicho propósito se utilizan como base dos archivos csv brindados por la empresa. En un primer lugar el archivo `events_up_to_01062018.csv` que contiene la información de los eventos realizados por un conjunto de usuarios hasta el 31 de mayo de 2018. En un segundo lugar el archivo `labels_training_set.csv` que determina si un conjunto de usuarios realizó o no una conversión desde el 01/06/2018 hasta el 15/06/2018.

Se propone como objetivo desarrollar una métrica a fin de evaluar los distintos resultados obtenidos. Se presentan en el informe las distintas decisiones adoptadas para elegir un resultado por sobre otro. Así mismo, se propuso utilizar distintos algoritmos de la librería sklearn para elegir el que modelice mejor.

Por otro lado se crearon distintos csv a fin de poder desarrollar de mejor manera el proceso de feature engineering y obtener un pre-procesamiento de los datos que permita encontrar los resultados más altos.

2 Investigación

Como se mencionó en la *Introducción*, se realizó un proceso de exploración de los datos para crear nuevos atributos y extraer nuevos features. De esta manera se crearon distintos csv con información extraída tanto de internet como de los propios datos para luego concatenar al set de datos final. Se procede a dar una breve explicación de la funcionalidad de cada uno de ellos. Muchos de ellos se encuentran detallados en el 'Notebook Anexo' del TP1[1].

- brands.csv

  Se agregó una columna al dataframe que detalla qué marca está involucarada en el evento del usuario.

- os.csv

  Se agregó una columna al dataframe que detalla qué sistema operativo está involucarada en el evento del usuario.

- browsers.csv

  Se agregó una columna al dataframe que detalla qué explorador de internet se accede al sitio.

- sessions.csv

  Se agregó el concepto de sesión, que se define como la agrupación de una serie de eventos por usuario, los cuales están todos con menos de 30 minutos de inactividad entre el actual y el anterior. Esto fue fijado con un criterio arbitrario a fin de poder discretizar el tiempo y definir este concepto.

- prices.csv

[1] https://fdelmazo.github.io/7506-Datos/TP1/TP1.html

Se agregó una columna al dataframe que indica el precio del producto involucrado en el evento del usuario. Para ello se extrajeron los precios de la página de Trocafone[2] considerando el sku, el modelo, el color, la capacidad de almacenamiento y la condición.

3 Features buscadas

4 Algoritmos utilizados

5 Desarrollo

6 Resultados obtenidos

7 Conclusiones

[2] https://www.trocafone.com/