Universidad de Costa RIca
PF-3115: Computational and statistical techniques of Machine Learning
Reporte de Laboratorio #5
Osvaldo Ureña A55783
María José Cubero B22148

Resumen

En este trabajo se propone una nueva técnica de clasificación y se compara con otra técnicas populares como la Bayesiana, Redes Neuronales, Árboles de decisión y Máquinas de Soporte Vectorial. Se aplica esta técnica para optimizar la predicción de campañas de marketing utilizando los datos de una entidad bancaria en Portugal.

Preguntas

Make sure to answer the following questions.

Regarding Data:

1. What is the source for the data used by your article?

Los datos están relacionados con campaña de marketing de una institución bancaria de Portugal y se tomó de The UCI Machine Learning Repository.

2. What biases does this data source have?

Se toman los datos de varios años, incluido el año 2008 donde se presentó una crisis económica. Además, los datos fueron divididos en el training data (4 años) y el test data (1 año), pero los resultados se pueden ver afectados por el año en específico que se seleccionó para el test data o los que se seleccionaron en el training data.

Por otro lado, el dataset está desbalanceado, ya que únicamente el 11.26% de los casos representan casos de éxito, lo cual puede afectar el entrenamiento.

3. Would there be a better way to collect the data?

Sería mejor poder solicitar los datos a la fuente primaria (el banco). Sin embargo, se considera difícil que ellos brinden este tipo de información por el tipo de entidad.

4. Do you think the obtained results generalize to the target population (and what is the target population)?

El público meta serían los clientes del banco de Portugal que aceptarían el producto bancario. Se considera que la generalización de los resultados puede verse afectada por dos factores:

- Primero, la economía es muy cambiante. Puede que el año en el que se quieran aplicar los resultados se esté pasando por una recesión o expansión económica, lo cual implicaría que la predicción no tome en cuenta la situación presente. Además, puede que para el entrenamiento se incluyeran años que presenten situaciones atípicos, como la crisis económica del 2008.
- Segundo, la forma en que se tratan los valores faltantes de las variables categóricas pueden afectar los resultados. Esto debido a que se menciona que hay bastantes valores faltantes, y lo que se hace es tomar el promedio, pero el promedio no necesariamente refleja la realidad.

Regarding Use:

1. What is the stated use/benefit of the work?

Proponer un nueva técnica de clasificación para mejorar la predicción de las campañas de marketing.

2. What other potential use-cases could the work be applied to? Which of them could be ethically problematic?

Usos potenciales: Utilizar esta nueva técnica de clasificación para otros tipos de trabajos. Como por ejemplo, compañías de seguros.

Problemas éticos: No mencionarle a los usuarios de forma explícita para qué fines se están utilizando sus datos. Además, para el caso de las compañías de seguros, podría tener varios problemas éticos al negarle a una persona el acceso a un seguro y servicios médicos solo por su historial.

3. Do the authors consider the potential for misuse? How do they propose to address it?

No se considera en ningún momento el mal uso que se le podría dar a esta técnica de clasificación.

Regarding Reproducibility:

1. Is the data and/or implementation freely available?

Los datos sí son de acceso público, pero la implementación no es de acceso público. De hecho, los autores no mencionan en qué software llevaron a cabo la ejecución de los procesos analizados en el paper, no se menciona el lenguaje utilizado ni si se hizo uso de alguna librería. Además, no detallan la forma automatizada en la que sacaron los valores más significativos, ya que aparentemente se utiliza un software para obtenerlos.

2. Could you (with the resources you have) reproduce the results?

Por los motivos mencionados anteriormente, no se puede reproducir el trabajo.

Referencias

Koumetio Tekouabou, Cédric Stéphane & Cherif, Walid & Hassan, Silkan. (2018). Optimizing the prediction of telemarketing target calls by a classification technique. 1-6. 10.1109/WINCOM.2018.8629675.