Predicción del éxito en campañas de mercadeo

Maria José Cubero
Osvaldo Ureña
June 30, 2020

Abstract

En el presente trabajo se pretende realizar predicciones sobre la respuesta de un individuo ante una campaña de mercadeo, utilizando herramientas de aprendizaje de máquinas. Esto nace con el fin de conocer cuáles personas podrían ser mejores cadidatos a la hora de ofrecerles un producto. Para lograrlo, se plantea utilizar estrategias como regresión logística, árboles de decisión y k vecinos más cercanos.

1 Introducción

Las campañas de marketing o venta de prodcutos son una estrategia muy común utilizada por las empresas para mejorar sus negocios.[1] Los autores antes citados también indican que la tecnología permite repensar el marketing enfocandose en lo que se conoce como "life time value" a través de evaluar la información disponible de los clientes y de esa forma poder desarrollar una relación con los clientes más alineada con las demandas del negocio.

Para las empresas que tienen grandes cantidades de clientes y que además le pueden ofrecer productos cruzados a sus clientes es de suma importancia poder realizar campañas eficientes en términos tanto económicos como de tiempo. Por ejemplo, un banco que tiene clientes con tarjetas ya sea de débito o crédito e intenta a ofrecer a esos clientes un seguro sobre su tarjeta, es más eficiente preparar campañas que abarquen clientes que podrían responder postivimante a la campaña y no llamar aquellos clientes que podrían responder negativamente, ahí es donde nace la importancia de poder utilizar modelos de clasificación que faciliten precisamente la elección de clientes para las campañas.

El problema que pretende resolver la investigación es **¿Qué técnica predice mejor el resultado de una campaña de telemercadeo para cada individuo?** Para responder lo anterior, uno de los primeros objetivos que nos planteamos es comparar varias técnicas como regresión logística, árboles de desición y k vecinos más cercanos. De modo que, se comparen las medidas de ajuste de las técnicas para identificar, en este caso concreto, cuál técnica podría ser más eficiente a la hora de elegir las personas meta para la campaña. Ahora teniendo en cuenta que las empresas ahora cuentan con una gran cantidad de información sobre sus clientes un segundo objetivo que nos planteamos es hacer una comparación en términos de eficiencia computacional de cada técnica.

2 Estado del arte

Desde hace ya varios años se ha venido hablando de la utilización de una serie de técnicas que se les ha llamado mineria de datos para mejorar actividades en los negocios, especialmente aquellos que por su naturaleza recolectan y manejan enormes cantidades de datos.[2] Minería de datos es la exploración y análisis de grandes cantidades de datos para descubrir patrones y reglas, el objetivo de la minería de datos es permitir a las corporaciones mejorar su marketing, ventas y operaciones de soporte al cliente a través de un mejor entendimiento de los clientes (Radharkrishnan y otros, 2013). Según los autores lo anterior se logra mediante seis tareas principalmente, que son: Calsificación, estimación, predicción, grupos de afinidad, clustering y perfilación. En este paper los autores hacen referencia al uso de árboles de decisión para clasificar aquellas personas que serían los mejores prospectos como clientes.

En el paper "Data Mining For Marketing"[3] los autores discuten como el proceso de marketing toma en cuenta mucha planeación e investigación y de que forma se pueden utilizar técnicas de data mining para hacer más eficiente y efectivo el proceso, dentro de las técnicas que mencionan para llevar a cabo una buena

[1] Moro, Cortez y Rita (2014)
[2] Radhakrishnan, Shineraj y Muhammed, 2013
[3] Mushtaq y Kanth, 2015

clasificación en aras de mejorar el proceso de marketing estan: Regresión Logistica, K vecinos más cercanos, Naive Bayes y árboles de desición.

Por otro lado los autores Basser e Imran en su paper "Use of Data Mining in Banking" resaltan la utilización de técnicas y algoritmos de data mining como clasificación, asociación, clustering, predicción y patrones secuenciales, explican también que una de las areas donde más se utilizan estas técnicas es en el área de marketing debido a que estas técnicas ayudan a determinar la conducta de los clientes referente a un producto, precios y canales de distribución, lo que puede dar paso a mejorar los canales y formas de promocionar un producto nuevo, o mejorar la calidad de servicios y productos ya existentes lo que se traduce en una mejora en productividad.

En el paper "Application of Data Mining in Banking Sector" se discute también la importancia de aplicar técnicas de data mining en el negocio bancario, resaltando su aplicación en Marketing, por ejemplo en la utilización de campañas para robar clientes a la competencia algo que puede ser posible gracias al análisis de los datos de los clientes esto para identificar factores que pueden afectar la demanda de los clientes en el pasado y las necesidades futuras; también las técnicas de data mining se pueden usar para volver más eficases y eficientes las estrategias orientadas al cliente y con que disposición un cliente podría aceptar nuevos productos, y un punto muy importante es como las técnicas de data mining pueden mejorar la eficiencia de la estrategia de fuerza de venta para un grupo determinado.

Existen dos papers relevantes el primero de ellos "A data-driven approach to predict the success of bank telemarketing" donde los autores utilizan una base de datos similar a la que nosotros vamos a utilizar, y utilizan cuatro métodos, Regresión Lógistica, Redes Neuronales, Árboles de Decisión y Support Vector Machine, esto con el objetivo de predecir si un cliente aceptaria un certificado de déposito a plazo bajo una campaña de telemercadeo, los autores concluyen que el método que da mejor resultado es el de Redes Neuronales. Algunos de los autores del paper anetriormente citado escribieron previamente un paper títulado " Enhancing Bank Direct Marketing through Data Mining" donde utilizaron parte de la base de datos que nosostros vamos a utilizar para predicir si un cliente aceptaría un certificado de déposito a plazo bajo una campaña de telemercadeo, pero esta vez solamente con un tipo de modelo Suppor Vector Machine, los autores conluyen que tuvieron muy buenos resultados.

Todas las fuentes consultadas concuerdan en la importancia de las metodólogias de clasificación, predicción y clustering en los negocios y particularmente como las técnicas de data mining pueden ayudar en la eficiencia de campañas de mercadeo así como en la elaboración de las mismas.

3 Materiales y Métodos

3.1 Materiales

Para realizar este trabajo se utilizará una base de datos disponible en el sitio web de "UCI Machine Learning Repository."[4] Dicha base fue utilizada para elaborar dos papers, los cuales serán de importancia para comparar la utilidad de las técnicas que se desarrollarán. Esta base cuenta con 41188 observaciones y 17 columnas, las cuales son:

- **age**: La edad del individuo (variable numérica).
- **marital**: Estado civil de la persona (variable categórica).
- **job**: Tipo de empleo de la persona (variable categórica).
- **education**: Educación del individuo (variable ordinal/categórica).
- **default**: Indica si la persona tiene un crédito en atraso (categórica).
- **balance**: Saldo de la persona (numérica).
- **housing**: Indica si la persona tiene un crédito hipotecario (categórica).
- **loan**: Indica si la persona tiene un crédito personal (categórica).

[4] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

- **contact**: Tipo de contacto ya sea por teléfono fijo o por celular (categórica).
- **day**: Día en que se contactó la persona(categórica).
- **month**: Mes en que se contacto la persona(categórica).
- **duration**: Duración del contacto en segundos (numérica).
- **campaign**: Número de veces que fue contactada la persona en la campaña (numérica).
- **pdays**: Número de días que han pasado desde que el individuo fue contactado en otra campaña (numérica).
- **previous**: Número veces que fue contactada la persona en campañas anteriores (numérica).
- **poutcome**: Resultado en la campaña anterior (categórica).
- **y**: Variable a predecir. Se refiere a si la persona acepta o rechaza el producto (categórica).

3.2 Métodos de clasificación

3.2.1 Regresión lógistica

Dentro de los métodos que se planean utilizar esta la regresión lógistica[5], que modela la siguiente probabilidad $p(X) = Pr(Y = 1|X)$ intentando usar un modelo de regresión lineal. Para representar esa probabilidad obtenemos $p(X) = \beta_0 + \beta_1 X$, ahora con la regresión logistica modelamos la relación anterior de la siguiente forma

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

manipulando algebráicamente la expresión obtenemos

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

y aplicando logaritmo a ambos lados de la ecuación resulta

$$log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X$$

el lado izquierdo de la ecuación se le llama log-odds o logit. Este método es adecuado para solucionar el problema de investigación, puesto que dado un set de entrenamiento con una matrix de "X" características y un vector de respuestas dicotomicas "Y", queremos clasificar la respuesta Y para una serie de clientes que presentan las misma matriz de características "X".

3.2.2 Árboles de desición

Otro método a utilizar es el de árboles de desición[6], en este caso supongamos que tenemos una partición de M regiones $R_1, R_2..., R_M$ y modelamos la respuesta de una constante c_m en cada región.

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

Adoptando el criterio de minimización de la suma de cuadrados $\sum(y_i - f(x_i))^2$, entonces el mejor $\hat{c}_m$ es el promedio de y_i en la región R_m

$$\hat{c}_m = ave(y_i|x_i \in R_m)$$

El algoritmo procede empezando por todo el set de datos considerando una variable j para hacer la división y un punto de división s, y define el par de medios planos.

$$R_1(j, s) = X|X_j \leq s \quad y \quad R_2(j, s) = X|X_j > s$$

[5]Información tomada del libro An introduction to statistical learning with applications in R
[6]Informacón tomada del libro The elements of statisticals learning. Data mining, inference and prediction.

Entonces se busca dividir la variable j y el punto de división s que resuelva

$$min_{j,s}\left[min_{c1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+min_{c2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

Este método fue utilizado en los papers que utilizaron la base de datos que nosotros utilizaremos.

3.2.3 K vecinos más cercanos (KNN)

La otra técnica a utilizar es la de K vecinos más cercanos (KNN)[7]. Los métodos de k vecinos más cercanos toman las observaciones del set de entrenamiento cercanas en un espacio de entrada x para formar $\hat{y}$, específicamente de la siguiente manera

$$\hat{Y}(x)=\frac{1}{k}\sum_{x_i\,k(x)}y_i$$

donde $N_k(x)$ es el vecindario de x, definido por los k puntos cercanos en la muestra de entrenamiento. La cercanía implica una medida, por ejemplo la distancia euclidiana. También consideramos esta técnica apropiada dado que esperariamos que personas con características similares respondan de manera similar a una campaña.

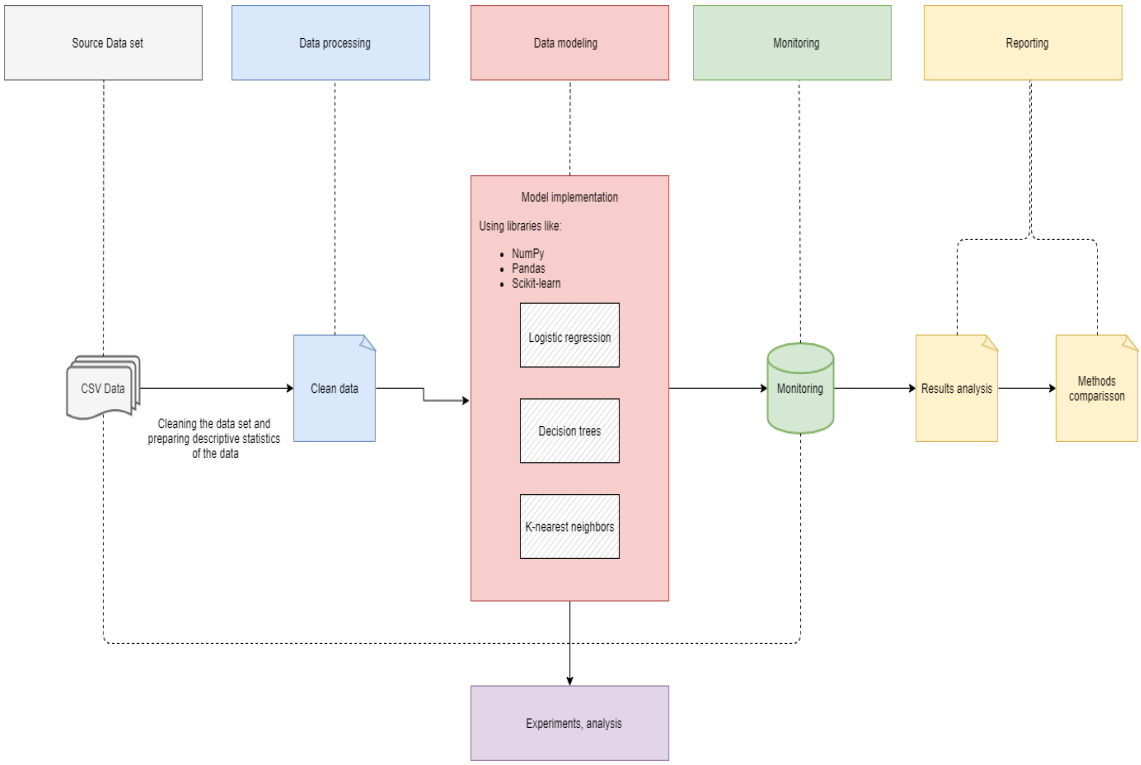
3.3 Flujo de trabajo

En el presente trabajo se realizará en 5 etapas, las cuales se presentan en el siguiente cuadro:

Cronograma		
Etapa	**Fecha**	**Trabajo a realizar**
Etapa I	05/05/2020	Flujo de trabajo y correcciones de primer avance.
Etapa II	12/05/2020 19/05/2020	Limpieza del conjunto de datos a utilizar y preparación de estadísticos descriptivos de los datos.
Etapa III	26/05/2020 02/06/2020 09/06/2020	Aplicar método de regresión logística. Aplicar método de árboles de desición. Aplicar método de K vecinos más cercanos.
Etapa VI	16/06/2020 23/06/2020	Análisis de resultados y comparación de métodos.
Etapa V	30/06/2020	Presentación.

[7]Informacón tomada del libro The elements of statisticals learning. Data mining, inference and prediction.

El flujo de trabajo es el que se encuentra a continuación:



4 Resultados

4.1 Análisis descriptivo

Una de las primeras cosas que nos interesa conocer es como se distribuyen las variables respecto a la aceptación o no del producto bancario. La primer característica pricipal que se puede observar es que el resultado no es balanceado, es decir, solamente un 12 % de las personas aceptaron el producto, el restante 88% no acepto el producto.

Observando por el tipo de trabajo en términos absolutos quienes más aceptaron el producto fueron las personas cuyo trabajo correspondían a management o technician, pero a nivel porcentual el producto fue más aceptado por personas cuyo trabajo era descrito como student y retired. Respecto al estado civil en términos absolutos el producto fue más aceptado por personas cuyo estado civil es married y en términos relativos tuvo más porcentaje de aceptación dentro de las personas single. Tomando en cuenta la educación el producto bancario fue mayormente aceptado por las personas con mayor grado de educación tertiary y secondary. Ahora observando si la persona tiene una hipoteca, un préstamo personal o había quedado en condición default en el pasado, se puede ver que la mayoría de aceptación fue por parte de quienes no tenían ninguna de las condiciones anteriores. Todo lo anterior se puede observar en el cuadro 1.

Cuadro 1
Estadísticos Descriptivos

Trabajo	NO	YES	Total	% Aceptación
student	527	216	743	29.07%
retired	1385	414	1799	23.01%
unemployed	885	163	1048	15.55%
management	6541	1051	7592	13.84%
unknown	205	29	234	12.39%
self-employed	1122	155	1277	12.14%
admin.	3628	497	4125	12.05%
technician	5444	657	6101	10.77%
services	3014	299	3313	9.03%
housemaid	927	91	1018	8.94%
entrepreneur	1084	94	1178	7.98%
blue-collar	7175	565	7740	7.30%
TOTAL	**31937**	**4231**	**36168**	**11.70%**
Estado Civil	**NO**	**YES**	**Total**	**% Aceptación**
single	8707	1522	10229	14.88%
divorced	3681	495	4176	11.85%
married	19549	2214	21763	10.17%
TOTAL	**31937**	**4231**	**36168**	**11.70%**
Educación	**NO**	**YES**	**Total**	**% Aceptación**
tertiary	9086	1609	10695	15.04%
unknown	1277	202	1479	13.66%
secondary	16593	1948	18541	10.51%
primary	4981	472	5453	8.66
TOTAL	**31937**	**4231**	**36168**	**11.70%**
Hipotecas	**NO**	**YES**	**Total**	**% Aceptación**
no	13376	2667	16043	16.62%
yes	18561	1564	20125	7.77%
TOTAL	**31937**	**4231**	**36168**	**11.70%**
Préstamo personal	**NO**	**YES**	**Total**	**% Aceptación**
no	26535	3847	30382	12.66%
yes	5402	384	5786	6.64%
TOTAL	**31937**	**4231**	**36168**	**11.70%**
Préstamo Default	**NO**	**YES**	**Total**	**% Aceptación**
no	31342	4189	35531	11.79%
yes	595	42	637	6.59%
TOTAL	**31937**	**4231**	**36168**	**11.70%**

Fuente: Elaboración Propia

4.2 Resultado de los modelos

Se corrieron 9 modelos, 3 modelos por cada técnica. La primer corrida fue con los datos tal como están originalmente en la base de datos. En una segunda corrida se agruparon las categorías de la variable job solamente. En la tercer corrida que produjo los mejores resultados se agruparon los datos de la variable job de la siguiente manera: (unemployed, housemaid, student, unknown) como others, (blue-collar y management) como bluecollar, (technician, admin, services) como tech-serv y (sel-employed y entrepeneur) como entrepeneur. Se reclasificó también la variable poutcome de la siguiente forma (unknown, failure, other) como others, la variable contact de la siguiente manera (unknow y telephone) como othercontact.

La nomenclatura utilizada es la siguiente los **verdaderos negativos** se refieren a la cantidad de observaciones que realmente rechazaron el producto y el modelo predijo que la persona iba a rechazar el producto, **falso negativo** se refiere a la cantidad de observaciones que realmente aceptaron el producto pero el modelo los clasificó como que rechazaron el producto, **falso positivo** se refiere a la cantidad de observaciones que realmente rechazaron el producto pero el modelo los clasificó como que aceptaron el producto, **verdadero positivo** se refiere a la cantidad de observaciones que realmente aceptaron el producto y el modelo los clasificó como que aceptaron el producto. **Accuracy** es el número de predicciones correctas divididas por el numero de predicciones, y **recall** son el número de verdaderos positivos para cada clase.

Cada una de las agrupaciones que se describen anteriormente obtuvo los siguientes resultados:

4.2.1 Sin agrupación

	Verdaderos negativos	Falso negativo	Falso positivo	Verdadero positivo	Accuracy	Recall (Yes)	Recall (No)
REGRESIÓN LOGÍSTICA	7290	720	175	338	0.90	0.32	0.98
ÁRBOLES DE DESICIÓN	7324	625	661	433	0.86	0.41	0.92
K VECINOS MÁS CERCANOS	7890	872	95	186	0.89	0.18	0.99

4.2.2 Agrupando la variable de job

	Verdaderos negativos	Falso negativo	Falso positivo	Verdadero positivo	Accuracy	Recall (Yes)	Recall (No)
REGRESIÓN LOGÍSTICA	7806	831	179	227	0.89	0.21	0.98
ÁRBOLES DE DESICIÓN	7289	616	696	442	0.85	0.42	0.91
K VECINOS MÁS CERCANOS	7963	1018	22	40	0.88	0.04	1

4.2.3 Agrupando la variable de job, poutcome y contact

	Verdaderos negativos	Falso negativo	Falso positivo	Verdadero positivo	Accuracy	Recall (Yes)	Recall (No)
REGRESIÓN LOGÍSTICA	7805	721	180	337	0.90	0.32	0.98
ÁRBOLES DE DESICIÓN	7290	625	695	433	0.85	0.41	0.91
K VECINOS MÁS CERCANOS	7871	832	114	226	0.90	0.21	0.99

5 Conclusiones

Según lo que se visualiza en la sección de resultados, se puede deducir que la cantidad de verdaderos positivos es muy baja con cualquiera de las técnicas en comparación con la cantidad de verdaderos negativos, que tiende a ser muy alta para este tipo de problemas. Sin embargo, estos resultados se pueden mejorar al agrupar algunas variables del dataset, como el job, poutcome y contact.

Por otro lado, al realizar una comparación entre una técnica de clasificación y otra se puede decir que cada una tiene sus ventajas y desventajas. Las que presentan el accuracy más alto son las de regresión logística y k vecinos más cercanos, pero de esas dos la que tiene el recall positivo más alto es la de regresión logística. Sin embargo, la que tiene una cantidad mayor de verdaderos positivos es la técnica de árboles de decisión.

Al hacer una comparación en términos de verdaderos positivos es claro que los árboles de decisión es la mejor técnica, pero al mismo tiempo se debe tomar en cuenta que es la que tiene mayor cantidad de falsos positivos, lo cual implica que se hagan más llamadas con una respuesta negativa por parte del cliente. Sin embargo, esta técnica es también la que presenta mejores resultados en los falsos negativos, lo cual permite descartar menos cantidad de clientes que en realidad iban a tener una respuesta positiva.

La técnica de k vecinos más cercanos es la que provee menos cantidad de verdaderos positivos y falsos positivos. Esto debido a que es la que presenta mayor cantidad de verdaderos negativos y falsos negativos. Lo anterior se puede interpretar como que se van a descartar una mayor cantidad de clientes que hubieran tenido una respuesta positiva y a pesar de que los falsos positicos son mucho menores que en las otras técnicas, la cantidad de verdaderos positivos también disminuye en gran cantidad.

No existe una técnica de clasificación mejor que la otra, ya que los resultados pueden parecer más favorecedores que otros dependiendo del objetivo de la entidad bancaria y de la capacidad de gasto de dicha entidad, esto por que, si el presupuesto para realizar la campaña es muy limitado al punto de llamar una cantidad de personas limitadas. Sin embargo, sí cabe destacar que en cuanto a eficiencia computacional la técnica de K vecinos más cercanos es la más lenta de todas, dura mucho más que las otras dos técnicas aplicadas.

6 Bibliografía

Basse, A. e Imran, K. (2012). Use of Data Mining in Banking. International Journal of Engineering Research as Applications. Volumen 2, Issue 2, March-April 2012

Bhambri, Vivek. (2011). Application of Data Mining in Banking Sector. International Journey of Computer Science and Technology. Volume 2, Issue 2, June 2011.

Hastie, T., Tibshiriani, R. y Friedman, J. (2009). The elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, New York

James, G., Witten, D., Hastie, T. y Tibshiriani, R. (2014). An Introduction to Statistical learning, with applications in R. Springer, New York

Mushtaq, A. y Kanth, H. (2015). Data Mining For Marketing. International Journal on Recent and Innovation Trends in Computing and Communication. Volume 3, issue 3, march 2015.

Moro, S. Cortez, P. Rita, P. (2014). A data-driven approach to predict the sucess of bank telemarketing. Decision Support Systems 62 (2014) 22–31

Moro, S. Cortez, P. Laureano, R. (2012). o Moro, Raul Laureano, Paulo Cortez, Enhancing bank direct marketing through data mining, Proceedings of the Forty-First International Conference of the European Marketing Academy, European Marketing Academy, 2012, pp. 1–8.

Radhakrishnan, B. Shineraj, G. Muhammed, K. (2013). Application of Data Mining In Marketing. International Journal of Computer Science and Network, Volume 2, Issue 5, October 2013.