UNIVERSIDAD DE ANTIOQUIA
1803

PREDECIR EL BENEFICIO DE MANTENER UN CLIENTE .

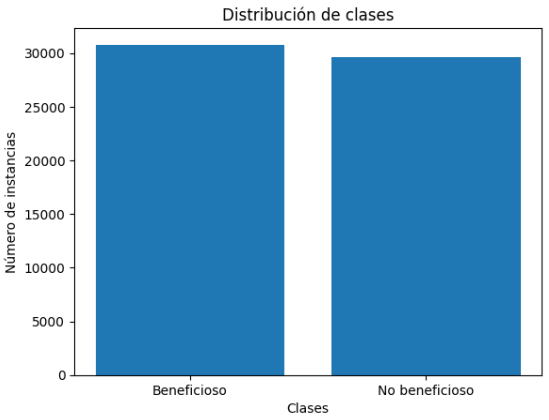
Santiago Bedoya Díaz, Juan Guillermo Buitrago Calle
Departamento de Ingeniería de Sistemas
Universidad De Antioquia, Colombia

RESUMEN

Nuestro problema de predicción consiste en que, por medio de aprendizaje, se entrene un modelo que nos diga el beneficio de mantener un cliente ligado al negocio o comercio de estudio, con el fin de predecir cuáles clientes son más importantes o más útiles en términos de ganancias y/o mantenibilidad financiera, por esto mismo el campo de aplicación de este modelo de predicción está enfocado a tiendas, supermercados o almacenes de cadena que tengan variedad de productos y clientes concurrentes, afiliados o de tipo cooperativa, en donde se ofrecen descuentos y demás beneficios a aquellos clientes fieles, y por lo mismo está la problemática de si esto genera beneficio real o no. Teniendo todo este panorama definido, se pretende que el modelo de predicción sea de clasificación, tomando a los clientes en rangos de costos, para saber si son beneficiosos/rentables para el negocio financieramente.

I. EXPERIMENTOS

En este proyecto se usa la base de datos ‘’Cost prediction on acquiring Customers’’ de kaggle https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart que contiene 60428 muestras, 39 variables de entrada y una de salida, de las cuales 17 son variables categóricas y 23 numéricas. Por medio de un gráfico se muestra la distribución de muestras por clase, y se puede ver que la base de datos no está desbalanceada.



Además tampoco cuenta con valores nulos.

```
#checking missing values
missing = df.isnull().sum()
missing = missing[missing > 0]
missing
```

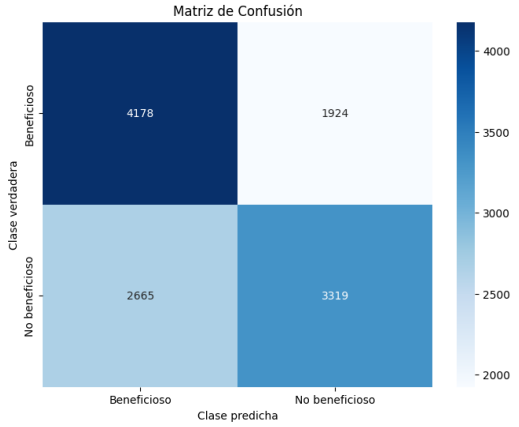
```
Series([], dtype: int64)
```

UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

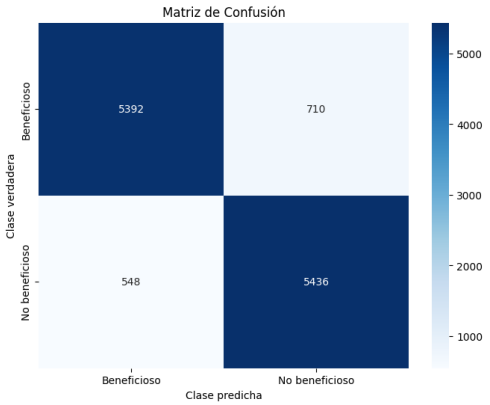
II. MEDIDAS DE DESEMPEÑO

Se usarán los modelos de Análisis discriminante Cuadrático, Gradient Boosting Tree, Redes Neuronales Artificiales y Máquinas de Soporte Vectorial. Los métodos de validación usados serán F1, accuracy, precisión, recall, support y matrices de confusión. En este caso se le dará mayor peso a la precisión. Se establecieron los conjuntos de entrenamiento y prueba, teniendo como el 80% de entrenamiento y 20% prueba. Para ello se usó el método ***train_test_split*** de la librería ScikitLearn**.**

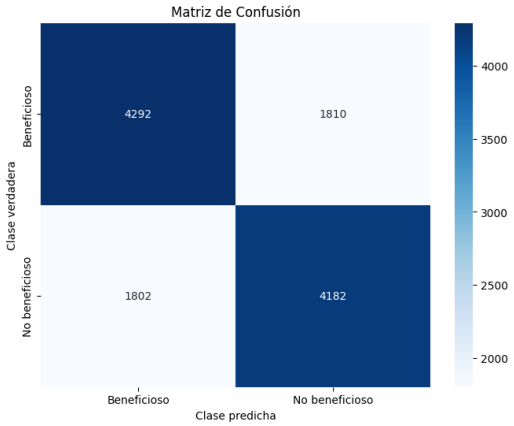
Análisis discriminante cuadrático:



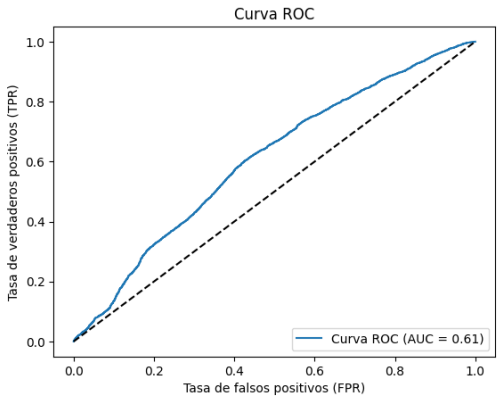
Gradient Boosting Tree:



Máquinas de Soporte Vectorial:



Redes Neuronales Artificiales:

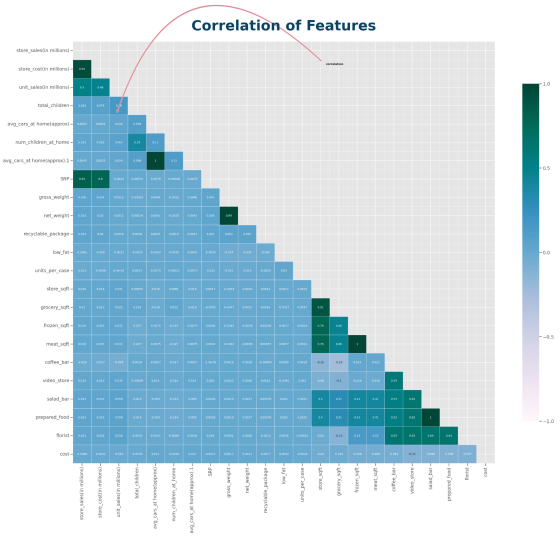


Nota: No se hace uso del método Ventana de Parzen porque es muy malo

computacionalmente y por eso en bases de datos grandes se prefieren otros.

III. SELECCIÓN Y EXTRACCIÓN DE CARACTERÍSTICAS

Se hace un análisis individual de todas las características a partir de la matriz de correlación para identificar variables que se correlacionan linealmente y aportan información redundante:



Se opta por escoger un umbral de +0.8 para identificar variables candidatas a ser eliminadas:

Variable 1	**Variable 2**	**Correlación**
store_sales(in millions)	store_cost(in millions)	0.95
store_sales(in millions)	SRP	0.83
store_cost(in millions)	SRP	0.8
avg_cars_at_home (aprox)	avg_cars_at_home (aprox).1	1
gross_weight	net_weight	0.99
store_sqft	grocery_sqft	0.91
frozen_sqft	meat_sqft	1
salad_bar	prepared_food	1

Usando el método de búsqueda secuencial descendente se tienen algunas observaciones: en el modelo que menos se demoró para encontrar las variables, que al eliminarlas el accuracy mejorará, fue en el Análisis Discriminante Cuadrático , pero su mejora fue de un 2% aproximadamente. En los modelos Gradient Boosting Tree y Máquinas de Soporte Vectorial se demoró mucho en dar resultados y al compararlos la mejora no era muy significante. Por ende según el método la única variable a eliminar sería store_sqft.

IV. COMPARACIÓN DE RESULTADOS CON LOS ARTÍCULOS CONSULTADOS

El artículo Customer Churn Prediction Based on Machine Learning usó algunos de los modelos que se usaron en este artículo, tales como Máquinas de Soporte Vectorial y Gradient Boosting Tree con métricas de validación como accuracy, precision, recall, F1-score, AUC, y matriz de confusión.

TABLE IV. THE SVM RESULTS

	Accuracy	Recall	Precision	F1-Score	AUC
Train	0.800	0.654	0.545	0.595	0.840
Test	0.800	0.611	0.524	0.565	0.830

TABLE VII THE GBDT RESULTS

	Rccuacy	Recall	Precision	F-measure	Auc
Train	0.830	0.729	0.587	0.650	0.890
Test	0.802	0.636	0.511	0.570	0.840

En este artículo los resultados son muy similares y allí indican que el mejor modelo se escogerá en una próxima investigación por medio de la optimización de Bayes. El artículo A review on Churn Prediction and Customer Segmentation using Machine Learning usó modelos como decision tree, multilayer perceptron con las mismas métricas de validación que en el anterior artículo y en este proyecto.

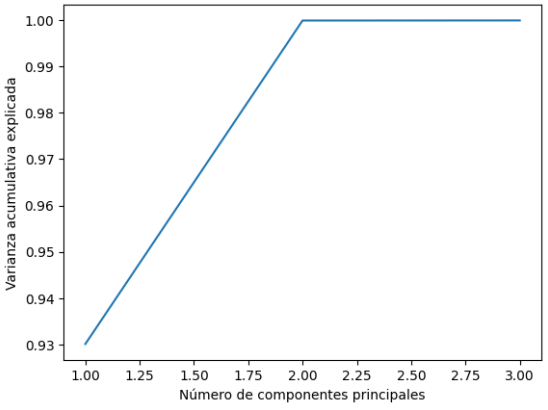
V. MODELOS Y VALIDACIÓN

Modelo	**Accuracy**	**Precision**	**Recall**	**F1**
QDA	0.6189	0: 0.61 1: 0.63	0: 0.68 1: 0.56	0: 0.64 1: 0.59
Gradient Boosting Tree	0.899	0: 0.90 1: 0.90	0: 0.90 1: 0.90	0: 0.90 1: 0.90
Support Vector Machine	0.70966	0: 0.72 1: 0.70	0: 0.71 1: 0.71	0: 0.71 1: 0.71
Multi-layer Perceptron	0.59159	0: 0.64 1: 0.57	0: 0.44 1: 0.75	0: 0.52 1: 0.64

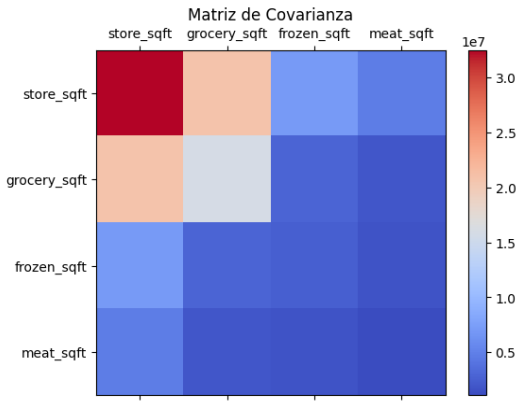
Dados estos datos de validación para cada uno de los modelos, podemos observar que aquel con mayor precisión fue el gradient boosting tree, con casi un 90% de precisión del modelo, tanto para detectar positivos como negativos, o mejor dicho, cada clase distinta. Por otra parte, nuestro peor modelo fue el MLP, que a pesar de probar con distinta cantidad de capas ocultas y neuronas, el 60% de precisión fue su mejor aproximación.

VII. ANÁLISIS CON PCA

Para el análisis con el método PCA para disminuir características, se usó la librería Scikit-Learn, con la cual podremos ajustar nuestros datos, ver la varianza de las características y por último sacar una matriz de covarianza luego, con el fin de analizar las variables resultantes.



Aquí podemos observar que luego de 2 componentes principales la varianza acumulativa tiende a 1.



Esta es la matriz de covarianza, acotada para que sólo enseñe las características más relacionadas entre sí, y saber cuáles podremos eliminar.
Al final, decidimos eliminar 'store_sqft' y 'grosery_sqtf' y correr de nuevo nuestros algoritmos.

IX. CONCLUSIÓN Y COMPARATIVA

Modelo	Precisión original	Precisión búsqueda secuencial	Precisión PCA	Mejora búsqueda	Mejora PCA
Quadratic Discriminant Analysis	0.6189	0.62295	0.6267	+0.405 %	+0.78 %
Gradient Boosting Tree	0.899	0.90435	0.8985	+0.535 %	-0.05 %
Support Vector Machine	0.70966	-	0.7082	-	-0.146 %
Multi-layer Perceptron	0.59159	-	0.5117	-	-7.989 %

En conclusión, nuestro mejor algoritmo en cuanto a la precisión para predecir si el cliente es o no beneficioso de mantener al negocio es el Gradient Boosting Tree, el cual nos da una precisión de aproximadamente 90% con los datos originales, luego le sigue Máquinas de Soporte Vectorial con 70%, y por último el Análisis discriminante Cuadrático con 61%.

En cuanto a las técnicas de reducción de datos, podemos ver que para búsqueda secuencial descendente tenemos mejoras casi mínimas, pero las hay, mientras que para el método PCA es contrario, tenemos disminución de precisión, lo cual nos dice que no es un método muy útil con el problema que estamos tratando. Cabe resaltar que la búsqueda secuencial para la cantidad de características que se estaban tratando tiene un costo computacional alto, por lo cual para algoritmos SVM y MLP se decidió no aplicar dicha técnica, ya que lo consideramos contraproducente.

Repositorio: https://github.com/Juanbc2/entregaFinalModelos2

UNIVERSIDAD
DE ANTIOQUIA
1803

Referencias

W. Yu and W. Weng, "Customer Churn Prediction Based on Machine Learning," 2022 4th International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Hamburg, Germany, 2022, pp. 870-878, doi: 10.1109/AIAM57466.2022.00176.

G. L. V. Prasad et al., "Machine Learning Based Cost prediction for Acquiring New Customers," 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2023, pp. 0866-0872, doi: 10.1109/CCWC57344.2023.10099189.

Baghla, S.; Gupta, G. Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-Commerce. *Microprocess. Microsyst.* **2022**, *94*, 104680.

A. Zadoo, T. Jagtap, N. Khule, A. Kedari and S. Khedkar, "A review on Churn Prediction and Customer Segmentation using Machine Learning," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 174-178, doi: 10.1109/COM-IT-CON54601.2022.9850924.