CLASES DESBALANCEADAS Y SELECCIÓN DE CARACTERÍSTICAS

Tópicos selectos de Cómputo

Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Victor Manuel Gómez Espinosa

1. MOTIVACIÓN

- Clasificación binaria: Bank Marketing Data Set (Machine Learning Repository)
- Objetivos:
- **Predecir si los clientes contratarán un producto** y **conocer cuáles son las características más importantes para la clasificación** ya que de esta forma se pueden **enfocar los recursos disponibles** en estas áreas para incrementar el éxito de futuras campañas (Moro et al. 2011).
- **¿a quién?, ¿cuándo?, ¿Cómo?**

2. METODOLOGÍA

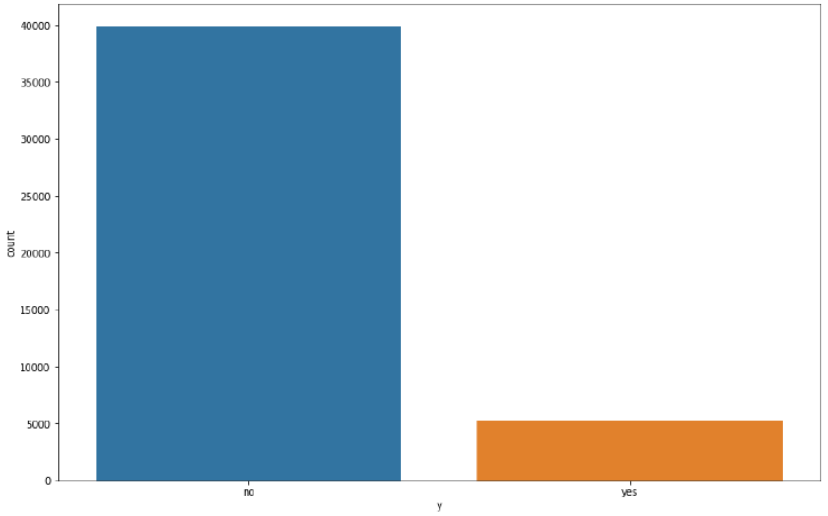

Figura 1: Clase de interés (*yes*) desbalanceada. Tasa de desbalance: 7.55.

- Métricas utilizadas: AUC, F1, Balanced Accuracy.
- Baseline: XGB
- Métodos para manejo de clases desbalanceadas:
- - Métodos a nivel de algoritmo: XGB y **SVM** con pesos.
- - Métodos a nivel de datos: Over-sampling (SMOTE), Under-sampling (ENN), métodos híbridos (SMOTE+ENN).
- - Métodos basados en preprocesamiento de datos y ensamble: RusBoost con AdaBoost como base.

2. METODOLOGÍA

$$Gain = \frac{1}{2}\left[\left(\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}\right)-\left(\frac{\left(G_L+G_R\right)^2}{\left(H_L+H_R\right)+\lambda}\right)\right]-\gamma$$

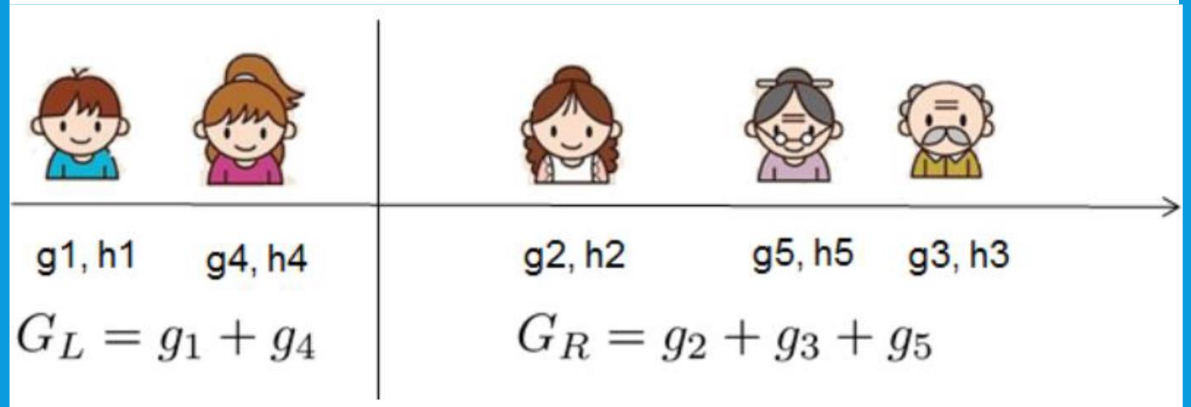

Figura 5: Ejemplo de corte que maximiza el *Gain* (Chen and Guestrin 2016)

- Selección de características bajo el criterio del *Gain* para el modelo XGBoost con pesos.

3. DISEÑO EXPERIMENTAL

- X(45211,16), y(45211), variables numéricas y categóricas.
- Test set 20%
- Se ajustaron los modelos mediante búsqueda aleatoria con validación cruzada 5-fold y métrica ROC-AUC.

- Hiperparametros:
- XGBoost: gamma: [0-1000], lambda:[1-1000], learning_rate: [0.1-0.9], subsample: [0.5-1], colsample_bytree:[0.5-1], scale_pos_weight : [1-100]
- SVM: C: [1-1000], gamma: [1e-3 – 1e3], kernel: rbf, class_weigth:balanced
- SMOTE: k_neighbors:[1-11]
- ENN: n_neighbors: [1-11]
- RUSBoost: learning_rate:[0.3-0.6], n_estimators:[100-300]

4. RESULTADOS

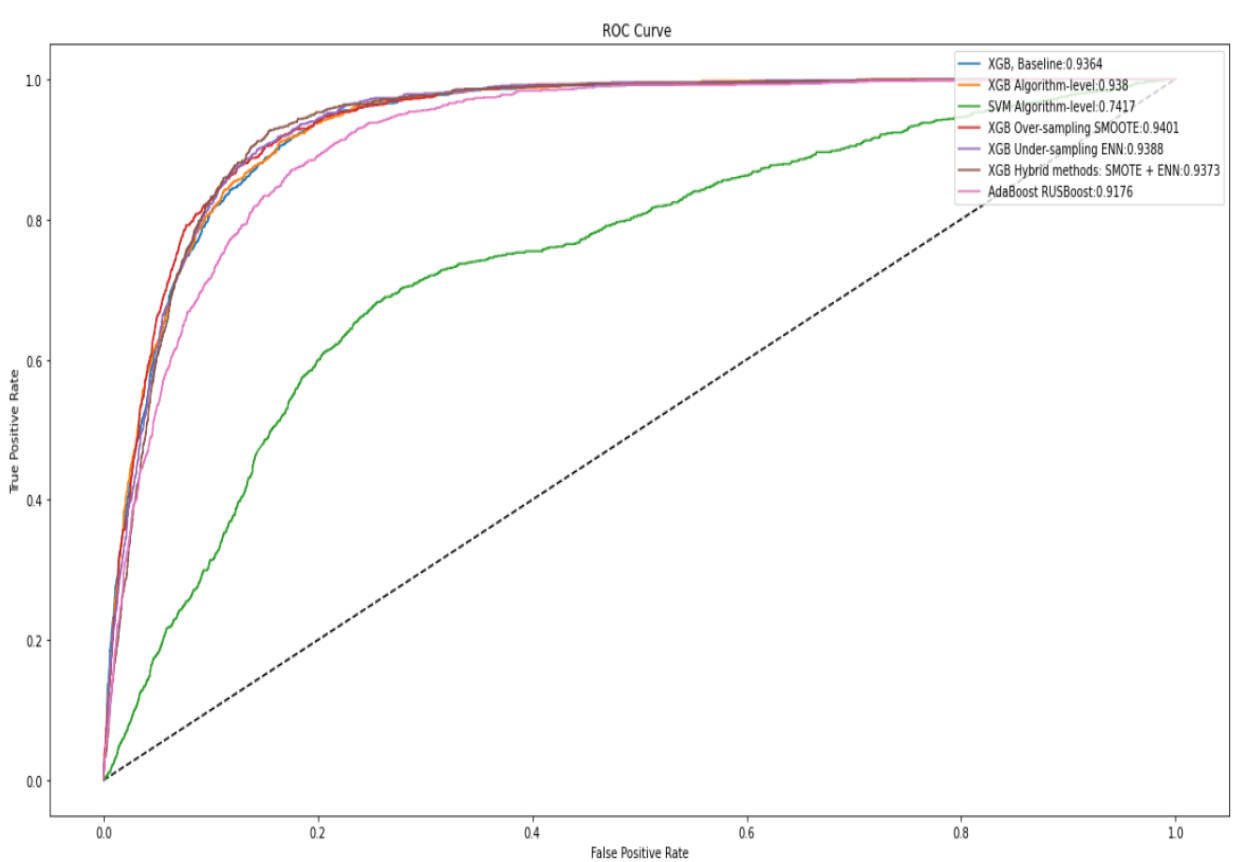

Figura 2: Resultados ROC Curve, AUC.

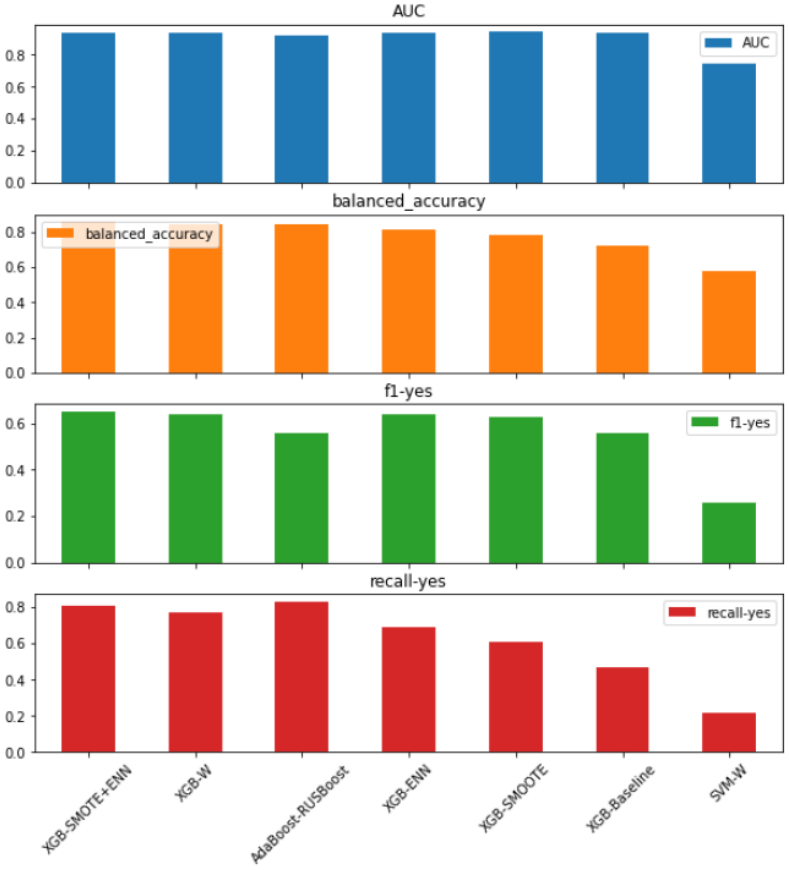

Figura 3: Resultados, métricas: AUC, F1-yes, Recall-yes.

4. RESULTADOS

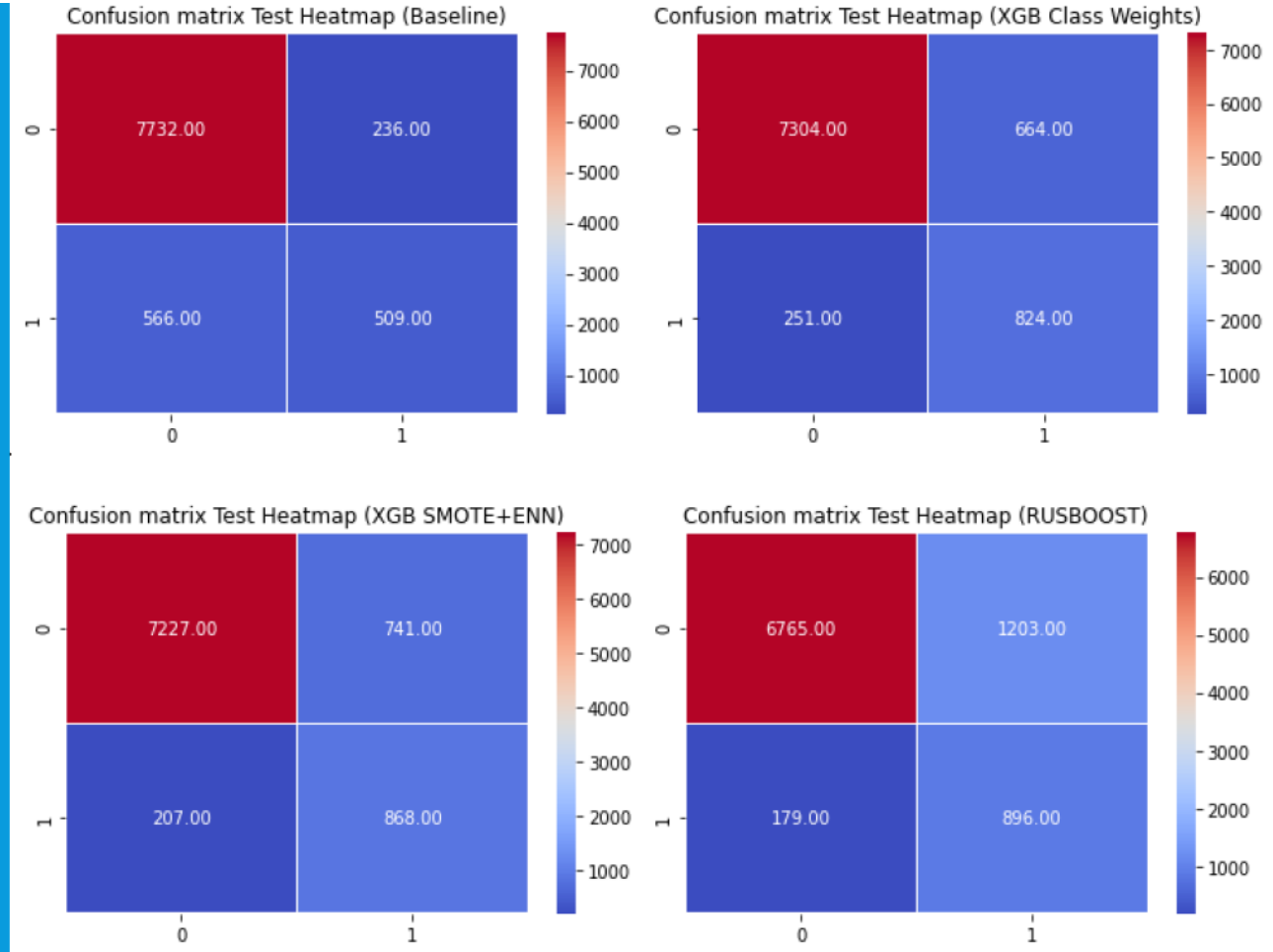

Figura 4: Matrices de confusión para el Baseline y los 3 mejores métodos.

4. RESULTADOS

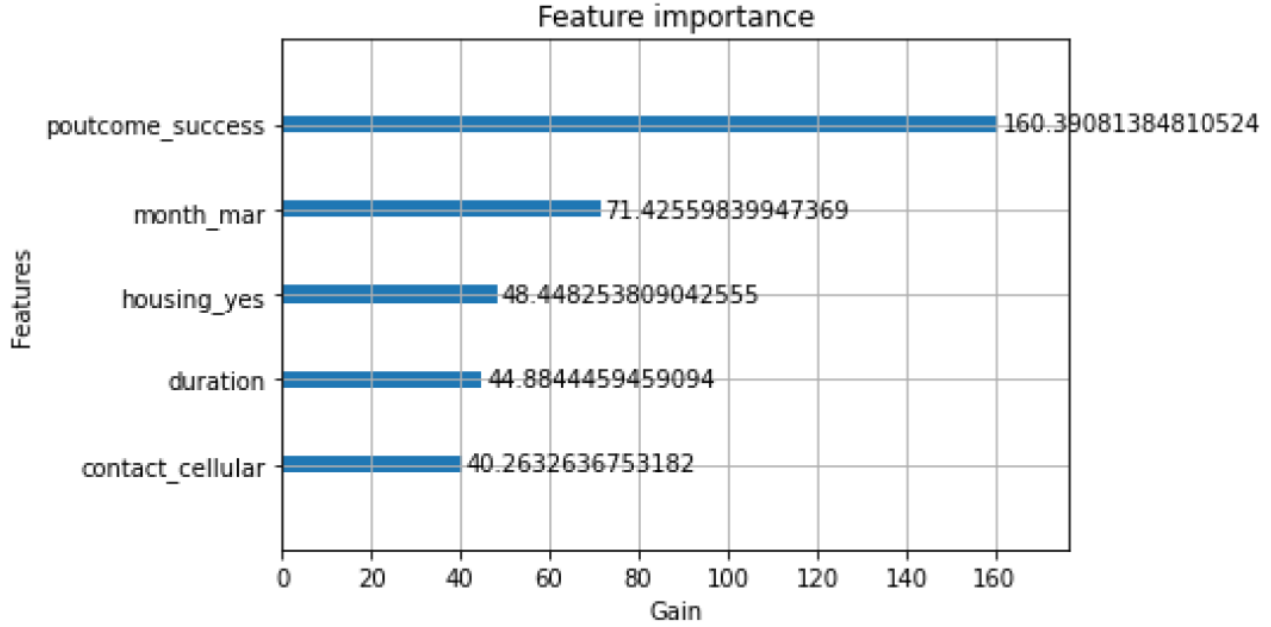

Figura 6: Características más importantes para el modelo XGBoost con pesos.

5. CONCLUSIONES

- Para este problema, **el método a nivel de algoritmo** fue el mejor, para el caso particular del modelo **XGBoost con pesos**, considerando el **AUC, Balanced Accuracy, F1** y su velocidad, sin embargo, como ya se mencionó, no sucedió lo mismo para el caso de máquinas de soporte vectorial.
- las **características más importantes** obtenidas del mejor modelo brindan gran información para los objetivos originales de este problema, ya que de alguna forma **permiten saber en qué segmento de usuarios concentrarse, cuando hacer la campaña y por qué medio contactar a los usuarios** para maximizar las probabilidades de éxito y disminuir quizá tiempo, dinero y recursos humanos.

6. REFERENCIAS

- Chen, T., and Guestrin, C. (2016), "XGBoost: A scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785.
- Moro, S., Laureano, R. M. S., and Cortez, P. (2011), "Using data mining for bank direct marketing: An application of the CRISP-DM methodology," *ESM 2011 - 2011 European Simulation and Modelling Conference: Modelling and Simulation 2011*, 117–121.