MÁSTER EN DATA SCIENCE

CURSO 2023-2024

MÓDULO: Introducción a la ciencia del dato.
PROFESOR: Abel González Durán

ALUMNO: Yonatán Eleuterio Rubio

Análisis De Datos Y Predicción De Bajas De Clientes

Introducción

Este informe documenta el proceso de análisis exploratorio de datos, preprocesamiento y modelado predictivo para un dataset de una entidad bancaria multinacional, centrado en predecir las bajas de los clientes. A través de este análisis, identificaremos las características más relevantes que influyen en la baja de los clientes y desarrollaremos modelos predictivos que ayuden a anticipar y mitigar las bajas.

Planteamiento

El objetivo principal es realizar un análisis exhaustivo del dataset proporcionado por el departamento de retención de la entidad bancaria. El análisis incluye:

1. **Carga y exploración inicial de los datos**.
2. **Análisis de la calidad del dato**.
3. **Preprocesamiento y limpieza de datos**.
4. **Análisis exploratorio de datos (EDA)**.
5. **Modelado predictivo**.
6. **Evaluación de los modelos**.

Ejecución y Resultados

1. Carga y Exploración Inicial de los Datos

Se cargan los datos y se revisan los primeros registros para entender la estructura del dataset y el tipo de información disponible.

2. Análisis de la Calidad del Dato

Se realiza una descripción estadística para obtener una visión general de las distribuciones y rangos de las variables. Además, se revisan los valores nulos y duplicados para evaluar la calidad del dato.

3. Preprocesamiento y Limpieza de Datos

Se imputan los valores nulos en las columnas numéricas usando la mediana y se eliminan los duplicados. Las variables categóricas se codifican utilizando LabelEncoder para convertirlas en formato numérico adecuado para el análisis y modelado.

4. Normalización de Datos

Se normalizan las variables numéricas para asegurar que todas las características tengan una escala comparable, lo que es esencial para el rendimiento de ciertos algoritmos de aprendizaje automático.

5. Análisis Exploratorio de Datos (EDA)

Se generan visualizaciones para entender la distribución con los datos de las diferentes categorías **('Geografia', 'Genero', 'TarjetaCredito', 'MiembroActivo', 'NumProductos')** y la **'Baja'** de clientes con graficas de barras en código de colores verde y rojo, marcando verde como lo positivo (**Numero de clientes que NO se dan de baja)** y marcando con rojo lo negativo **(Número de Bajas)** , creamos set de gráficos para visualizar en "boxplot" las diferentes categorías **('ScoreCredito', 'Edad', 'Antiguedad', 'Balance', 'SalarioEstimado')** cruzado con la **'Baja'** para ver los datos como se distribuyen con sus rangos en las diferentes categorías . También se calcula y visualiza una **matriz de correlación** para identificar relaciones entre las **Categorias**.

Por ultimo se crea una **"Pair Plot"** o **Gráfico de Pares** considerando los códigos categóricos antes creados para una visualización MACRO de las diferentes categorías correlacionadas y poder buscar tendencias dentro de las diferentes gráficas relacionadas.

6. Modelado Predictivo

Se divide el dataset en conjuntos de entrenamiento y prueba. El propósito del análisis fue predecir la baja de clientes en una institución financiera utilizando diversos modelos predictivos. Los modelos evaluados incluyen **Regresión Logística, Árbol de Decisión, Bosque Aleatorio y Gradient Boosting.** Los resultados de estos modelos se compararon en términos de **precisión, recall y f1-score** para determinar su eficacia en la predicción.

Resultados y Análisis

```
Logistic Regression Accuracy: 0.8166666666666667
              precision    recall  f1-score   support

           0       0.83      0.97      0.90      2416
           1       0.60      0.17      0.27       584

    accuracy                           0.82      3000
   macro avg       0.72      0.57      0.58      3000
weighted avg       0.79      0.82      0.77      3000

Decision Tree Accuracy: 0.7986666666666666
              precision    recall  f1-score   support

           0       0.88      0.87      0.87      2416
           1       0.48      0.52      0.50       584

    accuracy                           0.80      3000
   macro avg       0.68      0.69      0.69      3000
weighted avg       0.80      0.80      0.80      3000
```

```
 Random Forest Accuracy: 0.8673333333333333
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      2416
           1       0.77      0.46      0.57       584

    accuracy                           0.87      3000
   macro avg       0.82      0.71      0.75      3000
 weighted avg       0.86      0.87      0.85      3000

 Gradient Boosting Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.88      0.97      0.92      2416
           1       0.78      0.46      0.58       584

    accuracy                           0.87      3000
   macro avg       0.83      0.71      0.75      3000
 weighted avg       0.86      0.87      0.86      3000
```

A continuación, se presentan los resultados de los modelos predictivos evaluados:

1. **Logistic Regression**:
    - **Precisión**: 0.82
    - **Recall (macro avg)**: 0.57
    - **f1-score (macro avg)**: 0.58

Interpretación:

- La precisión general del modelo es del 82%, lo que indica que el modelo predice correctamente el 82% de los casos.
- La clase 0 (clientes que no se dan de baja) tiene una precisión y recall altos, con un f1-score de 0.90.
- La clase 1 (clientes que se dan de baja) tiene una baja precisión (0.60) y un muy bajo recall (0.17), lo que sugiere que el modelo no identifica correctamente a muchos clientes que se dan de baja.

2. **Decision Tree**:
    - **Precisión**: 0.80
    - **Recall (macro avg)**: 0.69
    - **f1-score (macro avg)**: 0.69

Interpretación:

- La precisión general es del 80%.
- La clase 0 sigue mostrando un alto rendimiento (f1-score de 0.87).
- La clase 1 mejora en comparación con la Regresión Logística, con una precisión de 0.48 y un recall de 0.51, lo que indica un mejor equilibrio entre precisión y recall para esta clase.

3. **Random Forest**:
    - **Precisión**: 0.87
    - **Recall (macro avg)**: 0.72
    - **f1-score (macro avg)**: 0.75

Interpretación:

- La precisión general del modelo es del 87%.

- La clase 0 tiene un rendimiento excelente (f1-score de 0.92).
- La clase 1 muestra una precisión de 0.76 y un recall de 0.47, indicando una mejora significativa en comparación con los modelos anteriores, aunque el recall aún puede mejorarse.

4. **Gradient Boosting**:
    - **Precisión**: 0.87
    - **Recall (macro avg)**: 0.71
    - **f1-score (macro avg)**: 0.75

Interpretación:

- La precisión general es del 87%, similar a la del modelo Random Forest.
- La clase 0 tiene un rendimiento excelente (f1-score de 0.92).
- La clase 1 muestra una precisión de 0.78 y un recall de 0.46, similar al rendimiento del modelo Random Forest, pero con una ligera mejora en la precisión.

Conclusiones

A partir de los resultados obtenidos, se pueden extraer varias conclusiones clave:

1. **Rendimiento General de los Modelos**:
    - Los modelos de Bosque Aleatorio y Gradient Boosting presentan las mejores precisiones generales (0.87), superando tanto a la Regresión Logística (0.82) como al Árbol de Decisión (0.80).
    - Estos modelos también muestran un buen equilibrio entre precisión y recall, especialmente para la clase 1, que es de particular interés.
2. **Desempeño en la Clase de Baja (Clase 1)**:
    - Aunque todos los modelos tienen un rendimiento significativamente mejor en la clase 0 (clientes que no se dan de baja), los modelos de Bosque Aleatorio y Gradient Boosting muestran una mejora notable en la predicción de la clase 1.
    - Aun así, el recall de la clase 1 sigue siendo un área de mejora, ya que la identificación de todos los clientes que se dan de baja es crucial para la acción preventiva.

3. **Aplicaciones Prácticas**:
    - Los modelos de Bosque Aleatorio y Gradient Boosting son los más recomendables para implementar estrategias predictivas en la institución financiera, debido a su alta precisión y mejor rendimiento en la identificación de clientes que se darán de baja.

En resumen, el uso de modelos de ensamble como Random Forest y Gradient Boosting ha demostrado ser eficaz en este contexto, proporcionando una herramienta poderosa para la predicción de la baja de clientes y permitiendo a la institución financiera tomar medidas proactivas basadas en estas predicciones.

Siguientes Pasos

1. **Optimización de Modelos**: Realizar una optimización de hiperparámetros para mejorar la precisión de los modelos.
2. **Análisis de Importancia de Variables**: Determinar las características más influyentes en la baja de clientes.
3. **Implementación en Producción**: Desplegar el modelo predictivo en un entorno de producción para la predicción en tiempo real.
4. **Estrategias de Retención**: Desarrollar estrategias basadas en las predicciones del modelo para retener a los clientes identificados con alto riesgo de baja.

Este proyecto proporciona un enfoque detallado para analizar y entender las bajas de clientes en una entidad bancaria, utilizando técnicas de ciencia de datos y machine learning. La correcta implementación y análisis permitirán al departamento de retención tomar decisiones informadas para mejorar la lealtad de los clientes y reducir las bajas.

Entrenamiento Modelo Predictivo

Se usan los modelos entrenados para predecir las probabilidades de que cada cliente se dé de baja. Esto permitirá identificar a los clientes con mayor riesgo.

Identificación de Clientes de Alto Riesgo

Se establece un umbral para clasificar a los clientes como de alto riesgo. Un umbral común es 0.5

Una vez que se tienen los clientes de alto riesgo, se puede analizar las características para entender mejor qué factores contribuyen a la posible baja.

Interpretación de Resultados

Probabilidades de Baja : rf_prob y gb_prob representan las probabilidades de baja predichas por los modelos de Random Forest y Gradient Boosting, respectivamente.

Clientes de Alto Riesgo: Los clientes con rf_prob o gb_prob superiores al umbral (0.5 en este caso) son considerados de alto riesgo.

Análisis de Características: Analizar las características de estos clientes puede revelar patrones o factores comunes que contribuyen a su riesgo de baja.