Challenge Analítica Avanzada e Inteligencia Artificial

Centro de Investigación Coppel

Elaborado por: Jhonathan Santacana

Descripción del problema:

El objetivo de este problema es desarrollar un modelo de contactabilidad que asocie una probabilidad de contacto a cada cliente en función de ciertos segmentos horarios. Para resolver este problema, se utilizará el archivo 20210513_Challenge_AA.csv, que contiene información sociodemográfica y transaccional de los clientes.

El proyecto se llevo a cabo a través de 3 etapas:

La primera etapa corresponde al proceso de análisis preliminar y ETL, implica cargar los datos del archivo CSV, seleccionar un horario específico, aplicar un filtro y trabajar con un dataframe más pequeño. Se realizan cálculos estadísticos básicos, se manejan los valores nulos y se analizan las correlaciones entre las variables. También se elaboran gráficos para visualizar la distribución de los datos y se sugieren posibles transformaciones para mejorar la capacidad predictiva de los modelos. El proyecto se centra en comprender la estructura de los datos, identificar patrones y relaciones, y preparar los datos para su análisis posterior.

La segunda parte del proyecto de machine learning se centra en la ingeniería de características. Se explora la extracción de información relevante de los datos existentes para mejorar la capacidad predictiva de los algoritmos de aprendizaje automático. Se utilizan diversas técnicas como la creación de nuevas características, la normalización, la codificación de variables categóricas, el manejo de valores faltantes y la reducción de dimensionalidad.

Se describe el proceso de selección de características, que puede realizarse mediante métodos de filtro (usando correlación, Valor de Información o Chi-cuadrado) o mediante métodos de envoltura (utilizando algoritmos de Machine Learning). También se detallan los métodos de la

clase "Preprocesamiento" que permiten aplicar transformaciones y codificaciones a las columnas del DataFrame, como la normalización, el escalado, la codificación One-Hot y el Weight of Evidence.

Se mencionan las características específicas de algunas variables, como la edad, el ingreso, el sexo y el estado civil, y se proponen transformaciones adecuadas para cada una de ellas. Además, se presentan los métodos de análisis de correlación, Information Value, Chi-cuadrado, PCA y la importancia de características de los árboles de decisión.

Por último, se discute la preparación del conjunto de entrenamiento y validación, incluyendo opciones como la división estándar, la validación cruzada, la validación estratificada y el muestreo balanceado, teniendo en cuenta el tamaño del conjunto de datos, la cantidad de muestras y la distribución de clases.El objetivo de esta parte del proyecto es optimizar las características del conjunto de datos y prepararlo adecuadamente para el modelado y la evaluación de los algoritmos de machine learning.

La tercera parte del proyecto consiste en la elaboración de modelos de clasificación y la evaluación de su rendimiento utilizando métricas como precisión, recall, F1-score y AUC-ROC. A continuación, se resumen las propuestas de modelos de clasificación presentadas en la tercera parte del proyecto:

1. Regresión Logística básica: Se utiliza como primer modelo para obtener una visión inicial del rendimiento de los datos. La Regresión Logística proporciona interpretación y comprensión de los coeficientes asociados a cada característica, es computacionalmente eficiente, menos propensa al sobreajuste y fácil de implementar.

2. Clasificador K Vecinos (KNN): Se considera debido al tamaño suficiente de la muestra y la estructura no lineal de los datos. KNN es un modelo simple y versátil que puede capturar patrones complejos sin hacer suposiciones sobre la forma de los datos.

3. RandomForestClassifier: Es un algoritmo de aprendizaje supervisado basado en un conjunto de árboles de decisión. Combina las predicciones de cada árbol individual para obtener una predicción final. RandomForestClassifier es conocido por su precisión, resistencia al sobreajuste, robustez y capacidad para proporcionar una medida de la importancia de las características.

4. SGDClassifier, que es una implementación del algoritmo de descenso de gradiente estocástico para problemas de clasificación lineal en scikit-learn. Se destaca su eficiencia y escalabilidad, así como su capacidad para ajustar hiperparámetros y su uso en conjuntos de datos grandes.

También se explica la importancia de las métricas de precisión, recall, F1-score y AUC-ROC para evaluar los algoritmos de clasificación. Estas métricas proporcionan una evaluación general del rendimiento del modelo en términos de su capacidad para capturar instancias positivas, evitar falsos positivos y falsos negativos, y distinguir entre las clases positiva y negativa.

Finalmente, se plantea la implementación de un modelo ganador utilizando el VotingClassifier de scikit-learn, que combina múltiples clasificadores individuales para tomar una decisión conjunta. Se mencionan las ventajas de utilizar el VotingClassifier, como la mejora de la precisión y la reducción de la variabilidad, y se destacan algunos parámetros que se pueden ajustar, como los clasificadores individuales, el tipo de votación y los pesos asignados a los clasificadores.

En resumen, el proyecto propone el uso de diferentes modelos de clasificación, la evaluación de su rendimiento mediante métricas y la implementación de un modelo ganador utilizando el VotingClassifier, el cual queda listo para predecir con futuros datos.