PROYECTO ANALÍTICA DE DATOS INTRODUCCIÓN A LA IA

Introducción a la IA

ELABORADO POR:

Elkin David Sánchez Velez
Jeffrey Schneider Rengifo Marín

Facultad de Ingeniería, Universidad de Antioquia.
Medellín, Colombia.
2023.

1. INTRODUCCIÓN

En el actual panorama empresarial, la capacidad de prever y comprender el comportamiento del consumidor es esencial para el éxito de cualquier estrategia de comercialización. En este contexto, la inteligencia artificial se presenta como una herramienta fundamental, brindando la capacidad de analizar vastos conjuntos de datos para extraer patrones y realizar predicciones precisas. El siguiente informe se centra en un desafío particular: la predicción de las ventas mensuales en línea de productos, específicamente programas de autoayuda en línea, luego de una campaña publicitaria inicial.

a) **Problema predictivo a resolver:** Predecir las ventas mensuales de un producto basado en características propias de un producto.
b) **Dataset:** El dataset usado es el siguiente: https://www.kaggle.com/competitions/online-sales/overview. El dataset posee 795 mil filas y 558 columnas, de las cuales consideramos relevantes las columnas OUTCOME_M1 a la OUTCOME_M12 que son las ventas mensuales online durante los primeros 12 meses después del lanzamiento del producto, además, tenemos dos datos temporales Date_1 que es el día en que inicia la campaña publicitaria principal y se lanza el producto al público, Date_2 es el número del día en que el producto fue anunciado y se lanzó la primera campaña publicitaria. El dataset también incluye datos categóricos Cat_x los cuales pueden estar marcados con 1 o 0 indicando si posee o no tal característica.
c) **Métrica de evaluación:** La métrica de evaluación es RMSLE (error logarítmico cuadrático medio) en las 12 columnas de predicción.

El RMSLE se calcula como:

$$e = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_1 + 1))^2}$$

2. EXPLORACIÓN Y PROCESAMIENTO DEL DATASET

Análisis Inicial de la Relación de las Columnas en el data set

En una fase inicial, se realizó un análisis de correlación entre las columnas del conjunto de datos utilizando la función .corr() y se generó un mapa de calor con sns.heatmap para visualizar las relaciones entre las variables.

Se aplicó un filtro a los valores de correlación, considerando solo aquellos mayores de 0.7, lo que nos permitió identificar las relaciones más fuertes. Esto proporcionó un punto de partida sólido para el análisis individual de cada columna.

Preprocesamiento de Datos

Se abordaron diversos aspectos del preprocesamiento de datos para garantizar la calidad y consistencia de los mismos. Se tomaron las siguientes medidas:

- **Limpieza de Valores Nulos:** Se identificaron y manejaron los valores nulos en las columnas "Quan_X" al rellenarlos con la media de la respectiva columna, evitando así alteraciones significativas en las métricas.

- **Normalización de Nombres de Columnas:** Se corrigieron errores en la nomenclatura de las columnas, como la confusión entre "Quant" y "Quan," asegurando la coherencia en la nomenclatura.

- **Corrección de Datos Erróneos en Columnas "Cat_X":** Se identificaron datos incongruentes en las columnas "Cat_X" y se igualaron a cero aquellos valores mayores o iguales a dos, ya que no se ajustaban a los parámetros iniciales.

- **Eliminación de Columnas Irrelevantes:** Las columnas Cat_X que contenían valores constantes en todas sus entradas se consideraron irrelevantes y se eliminaron del conjunto de datos.

- **Manejo de Fechas:** Se realizó una corrección en las fechas "Date_2," asegurando que los valores nulos sean iguales al menor de "Date_2," lo cual resulta coherente con la descripción de los campos.

3. DESARROLLO DE LA SOLUCIÓN

Análisis de la predicción

El proceso de implementación del modelo predictivo es crucial para evaluar su eficacia y comprender su capacidad para generalizar patrones a partir de los datos de entrenamiento. A continuación, se presenta un análisis detallado de las decisiones tomadas en la implementación y los resultados obtenidos.

Selección del Modelo

La elección del modelo es un paso fundamental en la construcción de un sistema predictivo. En este caso, se optó por utilizar el modelo de regresión lineal de la biblioteca scikit-learn, importando específicamente la clase LinearRegression. Dos parámetros clave se especificaron durante la importación:

- **fit_intercept=False:** Después de experimentar con diferentes configuraciones, se observó una optimización en el tiempo de ejecución al establecer este parámetro en falso. Al no ajustar una intersección, el modelo se ajusta directamente a través del origen, lo que puede ser adecuado en este contexto.
- **positive=True:** Dado que estamos tratando con ventas mensuales de productos, se asumió que las predicciones deben ser siempre valores positivos. Configurar este parámetro asegura que las predicciones no incluyan valores negativos, lo cual es coherente con la naturaleza del problema.

Función de Relación y Purga de Datos

Se definió una función de relación para evaluar la importancia de cada columna en el modelo. Esta función proporciona información sobre el peso de cada característica, permitiendo así una purga más efectiva de columnas menos relevantes. En particular, se decidió mantener solo aquellas columnas cuya relación con la variable objetivo supera el 10%. Este umbral puede ajustarse según las necesidades específicas del problema.

Métrica de Evaluación

La métrica de evaluación seleccionada es el RMSLE (error logarítmico cuadrático medio) en las 12 columnas de predicción. Esta métrica es adecuada para evaluar la precisión de las predicciones en un problema de regresión, proporcionando una penalización adicional para las discrepancias en valores más altos. La fórmula utilizada para el cálculo del RMSLE se presenta en la introducción.

División de Datos y Ajuste del Modelo

El conjunto de datos se dividió en dos partes: una para entrenar el modelo de regresión y otra para realizar la primera predicción y evaluar la calidad del ajuste. Durante este proceso, se aplicó un ciclo iterativo con el objetivo de encontrar un valor aceptable para la función de relación. Después de analizar diversas soluciones, se determinó que un valor de relación superior al 0.83 era óptimo para la optimización del tiempo de ejecución.

Evaluación del Modelo

Se realizaron predicciones tanto en la base de datos original como en la porción separada para evaluar la eficiencia del modelo. El error obtenido se midió mediante el RMSLE, y se observó que en algunas predicciones, el error era muy bajo, llegando en ocasiones a valores inferiores a 0.7 para algunas columnas.

4. RETOS DURANTE LA IMPLEMENTACIÓN DEL MODELO

Aquí se detallan algunos de los retos que surgieron durante este proceso:

1. **Manejo de la Gran Cantidad de Información:** En la fase inicial, la enormidad del conjunto de datos planteó un desafío significativo. La dificultad radicaba en analizar de manera precisa y significativa una gran cantidad de información. Este desafío requirió estrategias de filtrado y exploración efectivas para comprender adecuadamente las relaciones y patrones presentes en los datos.

Para abordar este reto, se aplicaron técnicas de análisis exploratorio de datos, como el uso de la función corr() y la visualización mediante un mapa de calor, lo que proporcionó una visión inicial de las relaciones entre las variables. Además, la función de relación creada posteriormente ayudó a focalizar el modelo en las características más relevantes.

2. Selección del Algoritmo de Predicción: Identificar el algoritmo de predicción más adecuado puede ser un proceso complejo y, en ocasiones, demorado. En este caso, la búsqueda del algoritmo óptimo llevó tiempo, y la asesoría del profesor fue esencial en esta tarea. Se exploraron diferentes opciones y configuraciones para encontrar el modelo que ofreciera el equilibrio adecuado entre rendimiento y precisión.

3. Confusión en los Datos de Variables Categóricas (Cat_X): La presencia de datos no conformes en las variables categóricas (Cat_X) añadió una capa de complejidad al proceso. La confusión radicó en la presencia de valores diferentes a 1 o 0 en estas columnas, desviándose de la esperada codificación binaria. Este desafío implicó una revisión exhaustiva y corrección de los datos incongruentes para garantizar la coherencia en el análisis.

4. Manejo de Predicciones de Ventas Negativas: Durante las fases de predicción, surgió un inconveniente crucial: la generación de predicciones de ventas negativas. Dado que las ventas no pueden ser negativas en el contexto de este problema, fue necesario encontrar una solución para limitar las predicciones a valores positivos. La configuración del modelo y la implementación de restricciones específicas, como el uso de positive=True, fueron estrategias clave para abordar este desafío y garantizar la coherencia con la naturaleza del problema.

Estos retos, aunque inicialmente desafiantes, han contribuido al fortalecimiento del proceso de implementación, proporcionando valiosas lecciones y mejorando la comprensión del conjunto de datos y del modelo predictivo. La superación de estos

obstáculos ha permitido obtener resultados prometedores y sienta las bases para futuras mejoras y refinamientos en el enfoque de predicción.

5. CONCLUSIÓN

La implementación de un modelo predictivo para prever las ventas mensuales en línea de programas de autoayuda ha sido un proceso desafiante pero revelador. La combinación de técnicas de análisis exploratorio, preprocesamiento de datos y selección de modelo ha proporcionado una sólida base para abordar la complejidad de un conjunto de datos extenso y diverso.

El análisis inicial reveló la importancia de comprender las relaciones entre las variables, destacando la necesidad de estrategias efectivas para manejar la gran cantidad de información presente en el dataset. La selección del modelo de regresión lineal, con parámetros cuidadosamente ajustados, demostró ser una elección acertada, respaldada por la asesoría del profesor durante la fase de exploración de algoritmos.

Los retos encontrados, como la confusión en los datos de variables categóricas y la generación de predicciones negativas, fueron abordados con enfoques específicos de preprocesamiento y configuración del modelo. Estos desafíos no solo demostraron la importancia de la atención meticulosa a los detalles, sino también la necesidad de ajustar y mejorar continuamente el enfoque para obtener resultados óptimos.

La métrica de evaluación, RMSLE, proporcionó una medida cuantitativa del rendimiento del modelo, permitiendo una evaluación precisa de la precisión de las predicciones. La obtención de errores bajos, incluso por debajo de 0.7 en algunas columnas, indica un rendimiento prometedor del modelo en términos de ajuste a los datos de entrenamiento.

En resumen, la implementación exitosa de este modelo predictivo subraya la importancia de la inteligencia artificial en el análisis de grandes conjuntos de datos para la toma de decisiones empresariales. Los desafíos superados durante el proceso han contribuido a un mayor entendimiento y capacidad de manejo de datos complejos. Este informe proporciona una base sólida para futuras iteraciones y mejoras en la implementación del modelo, respaldando así la continua aplicación de la inteligencia artificial en la comprensión y predicción del comportamiento del consumidor.