Adopción de técnicas de aprendizaje supervisado para predicción de categorías de productos en el sector minorista

Seminario

Entregable II

Maria del mar Ipia Guzmán
cc:1214726595
mdmguzman@hotmail.es

Especialización en Analítica y Ciencia de Datos
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Colombia

RESUMEN: Se analiza un conjunto de datos de ventas minoristas que contiene información sobre las transacciones de compras realizadas en 10 diferentes tiendas en Estambul, entre el 2021 hasta el 2023. Contiene información esencial acerca de las identificaciones de los clientes, edad, género, métodos de pago, categorías de productos, cantidad, precio, fechas de pedidos y nombres de las tiendas. Con estas características se pretende realizar un modelo de aprendizaje supervisado con el cual se pretende hacer una predicción de las categorías de compras en el mercado minorista.

PALABRAS CLAVE: Machine learning, clasificación, retail.

I. INTRODUCCION

El avance de la tecnología y la proliferación de grandes cantidades de datos han dado lugar a un creciente interés en el campo del aprendizaje automático y en las aplicaciones que puede tener este en diferentes sectores. Uno de ellos es el mercado minorista actual, donde la cantidad de datos disponibles es abrumadora, y en cual el uso del machine learning se ha convertido en un recurso invaluable para comprender y aprovechar la información generada por las transacciones de compra. En este informe, se pretende explorar sus aplicaciones en la clasificación de categorías de compras en el mercado minorista, por medio de metodologías de tareas de aprendizaje supervisado.

La clasificación precisa de las compras en diferentes categorías es fundamental para las empresas minoristas, ya que proporciona una visión profunda de los patrones de compra, el comportamiento del consumidor y las preferencias del mercado. Sin embargo, a medida que crece la cantidad de productos disponibles y se diversifican las categorías, se vuelve cada vez más difícil realizar esta clasificación de forma manual y eficiente.

Para ello, se utilizarán diversas técnicas y algoritmos de clasificación, como los modelos de clasificación RandomForest y Decision Tree Classifier que pueden analizar grandes volúmenes de datos para identificar patrones y realizar predicciones precisas. Se evaluará la precisión del modelo con el Accuracy_score. Además, se examinarán los beneficios de utilizar el machine learning en el mercado minorista tales como permitir a las empresas minoristas tomar decisiones basadas en datos, maximizar la eficiencia operativa y mantenerse competitivas en un entorno en constante cambio.

II. RECURSOS

Para la realización de este informe, se utilizaron herramientas tales como kaggle para la obtención de los datos, python y jupyter lab para el procesamiento y manipulación de los datos, y un repositorio de git hub en el cual se publicarán los resultados obtenidos.

III. ENTREGABLE I

A. Comprensión del problema de aprendizaje automático

1. Planteamiento del problema

El análisis de datos en el campo de la analítica y ciencia de datos desempeña un papel crucial en la comprensión de los patrones y tendencias dentro de un conjunto de datos. En este caso particular, se tiene acceso a un conjunto de datos de ventas minoristas que se presenta en un archivo CSV y contiene 99,457

registros y 10 columnas. Las columnas en el conjunto de datos son:

- "Date": La fecha en que se realizó la compra
- "Time": La hora a la que se realizó la compra
- "Total_amount": El importe total de la compra
- "Total_quantity": El número total de artículos comprados
- "Category": La categoría de los productos comprados
- "Card_type": El tipo de tarjeta de fidelización utilizada en la transacción
- "Payment_type": El método de pago utilizado en la transacción
- "Hour": La hora a la que se realizó la compra, separada de la columna "Time"

El objetivo principal de este proyecto de análisis es desarrollar un modelo de aprendizaje supervisado que pueda predecir las categorías de compras en el mercado minorista. Al aprovechar la gran cantidad de datos recopilados, se puede entrenar un modelo que pueda identificar patrones ocultos y relaciones entre las variables, y utilizar esta información para realizar predicciones precisas. Se utilizarán diversas técnicas y algoritmos de clasificación, como los modelos de clasificación RandomForest y Decision Tree Classifier.

El resultado de este análisis y la implementación exitosa del modelo de aprendizaje supervisado permitirán a las tiendas minoristas en Estambul mejorar su toma de decisiones estratégicas y optimizar sus operaciones. Al comprender mejor las preferencias de los clientes y las categorías de compras, las tiendas podrán adaptar su inventario, estrategias de marketing y promociones para satisfacer las necesidades de los consumidores de manera más efectiva, aumentando así su rentabilidad y competitividad en el mercado. En última instancia, este análisis contribuirá al crecimiento y desarrollo del sector minorista en Estambul.

2. Estado del arte

- Practical Data Science. (s.f.). How to Create a Naive Bayes Product Classification Model. Recuperado de https://practicaldatascience.co.uk/machine-learning/how-to-create-a-naive-bayes-product-classification-model

El proyecto se enfoca en la clasificación de productos en categorías correctas. Se utiliza un modelo de clasificación basado en Naive Bayes Multinomial y técnicas de Procesamiento del Lenguaje Natural (NLP) para predecir las categorías de productos a partir de sus nombres. Esto es relevante para mercados en línea y análisis de competencia. Se trabaja con un dataset de 35.311 nombres de productos, los cuales son asignados a 12.849 productos.

Como metodología, se utilizó vectorización de texto para aumentar la precisión del modelo y se trabajó con los hiperparámetros por defecto. Cómo métrica de evaluación se utiliza accuracy score y F1 score, dando como resultado 94.95% y 94.5% respectivamente.

- ILYA YATSYSHIN (2023). Customer Shopping dataset. Recuperado de https://www.kaggle.com/code/ilyai332/customer-shopping

El informe trabaja con el mismo data set empleado en este trabajo. Realiza una exploración exhaustiva de los datos y realiza transformaciones a las características, con el fin de obtener una mayor exactitud en el modelo.

A diferencia de este informe, el trabajo realiza una clasificación del precio de los productos, utilizando solo un modelo de random forest, con un hiperparámetro y la métrica utilizada fue el accuracy del 99.9%.

- Prakhar Gurawa (2022). Creating an E-Commerce Product Category Classifier using Deep Learning — Part 1. Recuperadode: https://prakhargurawa.medium.com/creating-an-e-commerce-product-category-classifier-using-deep-learning-part-1-36431a5fbc4e

En resumen, el objetivo del proyecto es desarrollar una API que pueda predecir las categorías de productos utilizando técnicas de aprendizaje automático y aprendizaje profundo. Se utilizará un conjunto de datos que contiene productos con sus categorías ya etiquetadas para entrenar los modelos de la API. La API resultante podrá clasificar nuevos productos en categorías específicas basándose en sus nombres y descripciones.

Hay un total de 51646 productos enumerados en el archivo product.json. Como cada instancia puede pertenecer a múltiples categorías, este tipo de problemas se conocen como problemas de clasificación de etiquetas múltiples, donde tenemos un conjunto de etiquetas objetivo.

Este trabajo busca aplicar una regresión logística a cada categoría del conjunto de datos. Para eso, utiliza la clase OneVsRestClassifier y con una métrica accuracy obtiene una precisión de 89.7%.

B. Entrenamiento y evaluación de modelos

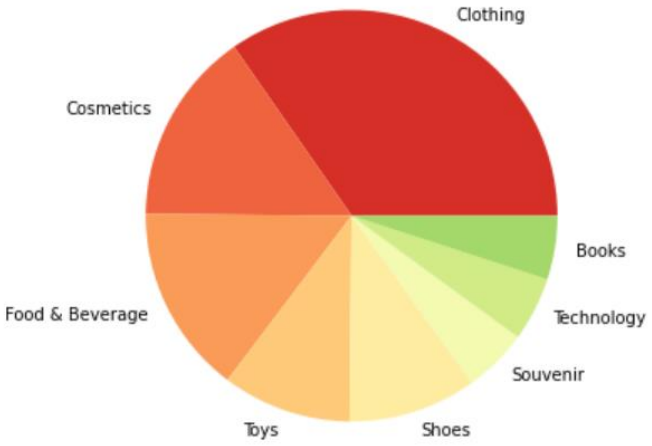
El conjunto de datos contiene 3 columnas numéricas y 7 columnas categóricas.

```
shopping_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99457 entries, 0 to 99456
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   invoice_no      99457 non-null  object
 1   customer_id     99457 non-null  object
 2   gender          99457 non-null  object
 3   age             99457 non-null  int64
 4   category        99457 non-null  object
 5   quantity        99457 non-null  int64
 6   price           99457 non-null  float64
 7   payment_method  99457 non-null  object
 8   invoice_date    99457 non-null  object
 9   shopping_mall   99457 non-null  object
dtypes: float64(1), int64(2), object(7)
memory usage: 7.6+ MB
```

No se realiza imputación, ya que se realiza el proceso de data clean al conjunto de datos y por medio del algoritmo LOF no se encuentran datos atípicos.

Es importante resaltar que, para la preparación de datos para el modelo, no se tiene en cuenta las columnas invoice_no, customer_id, invoice_date y shopping_mall, debido a que no aportan información relevante.

El objetivo principal de este trabajo es predecir las categorías de productos, las cuales se distribuyen de la siguiente forma:



Se realiza la agrupación de las categorías con menos participación en las transacciones para evitar que el modelo sólo pueda clasificar aquellas que tienen muchos registros, en este caso Clothing. Para ello, se agrupan las categorías Souvenir y Books y se elimina la categoría Technology debido a que no tiene datos suficientes para el modelo.
Para hacer uso de las variables categóricas en los modelos, se decide aplicar get dummies a las características gender y payment_method.

Se plantean dos modelos de clasificación: random forest y decision tree clasifier y por medio de la clase GridSearchCV, se busca encontrar cuales son los mejores hiperparámetros de los modelos. Como métrica de evaluación, se utiliza accuaracy score.

C. Resultados

Como resultado de los modelos utilizados, se obtiene por medio del ramdon forest un accuracy del 100% y el mismo valor para el modelo de decision tree clasifier. Debido a que los valores de exactitud son altos, se decide hacer un cross validation para descartar el sobreajuste. Como resultado de eso, se obtiene una precisión del 99.9%.

De acuerdo con la literatura consultada, los niveles de exactitud para los modelos de clasificación se encuentran por encima del 90%, y el tratamiento de los datos es similar a lo planteado en este informe.

El datalle y el procedimiento documentado se encuentra en el repositorio:

https://github.com/Mdm2016/Monografia-udea-cohorte-5/tree/main

D. Referencias

- Practical Data Science. (s.f.). How to Create a Naive Bayes Product Classification Model. Recuperado de https://practicaldatascience.co.uk/machine-learning/how-to-create-a-naive-bayes-product-classification-model
- ILYA YATSYSHIN (2023). Customer Shopping dataset. Recuperado de https://www.kaggle.com/code/ilyai332/customer-shopping
- Prakhar Gurawa (2022). Creating an E-Commerce Product Category Classifier using Deep Learning — Part 1.

- Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. hwrnaw. Recuperado de

  https://prakhargurawa.medium.com/creating-an-e-commerce-product-category-classifier-using-deep-learning-part-1-36431a5fbc4e