API 参考

发行版本 *v1.0*

lihetong

2025 年 12 月 26 日

内容目录:

欢迎来到 EcommerceSalesAnalysis 项目文档!

这是一个电商平台销售数据分析项目，包含数据加载、清洗、探索、可视化、特征工程、模型训练和评估等模块。

CHAPTER 1

介绍

这是一个电商平台销售数据分析系统，主要功能包括：

- 数据加载与预处理
- 数据探索与分析
- 数据可视化
- 特征工程
- 模型训练与评估

CHAPTER 2

模块参考

data_loader	
data_clean	
data_explore	
data_visualize	
feature_engineer	
model_train	
model_evaluate	
config	
main	

2.1 data_loader

Functions

data_loader(data, information)	
get_basic_info(df)	获取数据基本信息
load_raw_data()	加载原始用户数据 - 读取 CSV / 数据库 - 不做任何清洗与修改 - 保证“原始性” :return: 返回加载好的原始数据

data_loader.**data_loader**(*data*, *information*) → DataFrame

data_loader.**get_basic_info**(*df: DataFrame*) → dict

获取数据基本信息

:param df: 传入原始数据:return: 返回部分数据

data_loader.**load_raw_data**() → DataFrame

加载原始用户数据 - 读取 CSV / 数据库 - 不做任何清洗与修改 - 保证“原始性”:return: 返回加载好的原始数据

2.2 data_clean

Functions

clean_data(df, numeric_cols, categorical_cols)	统一保存:return:
handle_categorical_missing(df, categorical_cols)	类别型特征缺失值处理：填充为'Unknown'
handle_missing_values(df, numeric_cols)	数值型特征缺失值处理：- 使用 -1 进行占位填充 - 同时构造缺失指示变量:param df: 原始数据:param numeric_cols: 数值列:return:
mark_abnormal_values(df)	异常值标记，不修改原始取值
remove_duplicates(df_new)	查询是否有重复项, 去除重复项:return:

data_clean.**clean_data**(*df: DataFrame*, *numeric_cols: list*, *categorical_cols: list*) → DataFrame

统一保存:return:

data_clean.**handle_categorical_missing**(*df: DataFrame*, *categorical_cols: list*) → DataFrame

类别型特征缺失值处理：填充为'Unknown'

data_clean.**handle_missing_values**(*df: DataFrame*, *numeric_cols: list*) → DataFrame

数值型特征缺失值处理：- 使用 -1 进行占位填充 - 同时构造缺失指示变量:param df: 原始数据:param numeric_cols: 数值列:return:

data_clean.**mark_abnormal_values**(*df: DataFrame*) → DataFrame

异常值标记，不修改原始取值

data_clean.**remove_duplicates**(*df_new: DataFrame*) → DataFrame

查询是否有重复项, 去除重复项:return:

2.3 data_explore

Functions

categorical_by_lifecycle(df, col)	分析不同用户中，某类别特征的分布情况:param df: :param col: :return:
explore_categorical_features(df)	类别型特征分布:return: 返回用户各类别数量
explore_missing_values(df)	统计各字段缺失率:param df: :return:
explore_numeric_features(df)	数值型特征描述性统计:param df: 原始的数据:return: 返回数值型特征的描述性统计信息, 并转置结果（T）使特征作为行显示

data_explore.**categorical_by_lifecycle**(*df*, *col*)

分析不同用户中，某类别特征的分布情况:param df: :param col: :return:

data_explore.**explore_categorical_features**(*df: DataFrame*) → dict

类别型特征分布:return: 返回用户各类别数量

data_explore.**explore_missing_values**(*df: DataFrame*) → DataFrame

统计各字段缺失率:param df: :return:

data_explore.**explore_numeric_features**(*df: DataFrame*) → DataFrame

数值型特征描述性统计:param df: 原始的数据:return: 返回数值型特征的描述性统计信息, 并转置结果（T）使特征作为行显示

2.4 data_visualize

Functions

plot_box_by_category(df, cat_col, num_col)	类别-数值关系箱线图
plot_categorical_by_lifecycle(df, col)	
plot_categorical_distribution(df, col)	类别型特征分布
plot_correlation_heatmap(df)	数值特征相关性热力图
plot_numeric_distribution(df, col)	数值型特征分布

data_visualize.**plot_box_by_category**(*df: DataFrame*, *cat_col: str*, *num_col: str*)

类别-数值关系箱线图

data_visualize.**plot_categorical_by_lifecycle**(*df*, *col*)

data_visualize.**plot_categorical_distribution**(*df: DataFrame*, *col: str*)

类别型特征分布

data_visualize.**plot_correlation_heatmap**(*df: DataFrame*)

数值特征相关性热力图

data_visualize.**plot_numeric_distribution**(*df: DataFrame*, *col: str*)

数值型特征分布

2.5 feature_engineer

Functions

build_features(df, categorical_cols)	特征工程主入口
one_hot_encode(df, categorical_cols)	对类别型特征进行 One-Hot 编码只需要对类别列进行翻译

feature_engineer.**build_features**(*df: DataFrame*, *categorical_cols: list*) → DataFrame

特征工程主入口

feature_engineer.**one_hot_encode**(*df: DataFrame*, *categorical_cols: list*) → DataFrame

对类别型特征进行 One-Hot 编码只需要对类别列进行翻译

2.6 model_train

Functions

train_model(df, target_col[, test_size, ...])	模型训练模块:param df: 特征工程后的数据:param target_col: 目标列:return: 训练好的模型 + 测试集

model_train.**train_model**(*df: DataFrame*, *target_col: str*, *test_size: float = 0.2*, *random_state: int = 123*, *model_type: str = 'rf'*)

模型训练模块:param df: 特征工程后的数据:param target_col: 目标列:return: 训练好的模型 + 测试集

2.7 model_evaluate

Functions

evaluate_model(model, x_test, y_test)	模型评估函数

model_evaluate.**evaluate_model**(*model*, *x_test*, *y_test*)

模型评估函数

2.8 config

2.9 main

Functions

main()

main.**main**()

CHAPTER 3

Indices and tables

- genindex
- modindex
- search

CHAPTER 4

项目介绍

EcommerceSalesAnalysis 是一个电商平台销售数据分析项目。该项目旨在通过数据科学的方法对电商平台的销售数据进行深入分析，提供有价值的业务洞察。

4.1 功能特性

- **数据加载**: 支持多种格式的销售数据加载
- **数据清洗**: 自动识别和处理数据中的异常值、缺失值
- **数据探索**: 提供丰富的统计分析功能
- **数据可视化**: 生成直观的图表展示数据特征
- **特征工程**: 提取有用的特征用于模型训练
- **模型训练**: 使用机器学习算法预测销售趋势
- **模型评估**: 评估模型性能并提供可视化结果

4.2 技术栈

- Python 3.x
- Pandas - 数据处理
- NumPy - 数值计算
- Matplotlib/Seaborn - 数据可视化
- Scikit-learn - 机器学习
- Sphinx - 文档生成

Python 模块索引

索引

R

T