E-commerce Furniture Dataset 2024

Internship Project Report

1. Introduction

This project analyzes online furniture sales data collected from an e-commerce platform. The goal is to understand sales patterns, pricing effects, and shipping impact on product sales using data analysis and machine learning techniques.

2. Objective

- Analyze furniture sales trends
- Study the relationship between price and sales
- Predict number of units sold using machine learning models

3. Dataset Description

Dataset file name: **ecommerce furniture dataset 2024.csv**. It contains 2000 records with product titles, prices, sold units, and shipping tags.

4. Tools Used

Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, VS Code, GitHub

5. Data Preprocessing

Data cleaning included removing unnecessary columns, converting price to numeric format, handling categorical shipping tags, and preparing text data for modeling.

6. Exploratory Data Analysis

EDA was performed to visualize sales distribution, price trends, and the impact of shipping tags on number of units sold.

7. Machine Learning Model

Linear Regression and Random Forest Regressor models were trained to predict the number of items sold. Product titles were vectorized using TF-IDF.

8. Results & Evaluation

Model performance was evaluated using Mean Squared Error (MSE) and R^2 score. Random Forest performed better due to its ability to capture non-linear relationships.

9. Conclusion

This project demonstrates how data analytics and machine learning can be applied to e-commerce sales prediction problems. The insights gained can help in pricing and marketing decisions.

10. Code Appendix

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score

df = pd.read_csv("ecommerce furniture dataset 2024.csv")

df.drop(columns=['originalPrice'], inplace=True, errors='ignore')

df['price'] = df['price'].replace('[\$,]', '', regex=True).astype(float)

df['tagText'] = df['tagText'].apply(lambda x: x if x == 'Free shipping' else 'others')
df['tagText'] = df['tagText'].astype('category').cat.codes

tfidf = TfidfVectorizer(max_features=100)
title_tfidf = tfidf.fit_transform(df['productTitle'])
title_df = pd.DataFrame(title_tfidf.toarray(), columns=tfidf.get_feature_names_out())

df = pd.concat([df, title_df], axis=1)
df.drop('productTitle', axis=1, inplace=True)

X = df.drop('sold', axis=1)
y = df['sold']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

lr = LinearRegression()
rf = RandomForestRegressor(n_estimators=100, random_state=42)

lr.fit(X_train, y_train)
rf.fit(X_train, y_train)

y_pred_lr = lr.predict(X_test)
y_pred_rf = rf.predict(X_test)

print("LR MSE:", mean_squared_error(y_test, y_pred_lr))
print("RF MSE:", mean_squared_error(y_test, y_pred_rf))
```