Customer Segmentation and Sales Prediction Survey

Abstract

This project applies machine learning to tackle **Customer Segmentation** and **Sales Prediction**, enhancing marketing strategies and sales performance. Customers are grouped into segments using **K-Means Clustering**, while **Linear Regression** forecasts future spending, achieving a **MAE of 1.85k** and an **R² score of 0.82**.

A **Tkinter-based GUI** ensures ease of use, allowing non-technical users to load data, perform segmentation, and visualize results. The pipeline includes extensive preprocessing, efficient model training, and robust evaluation, resulting in a scalable and interpretable solution that supports strategic decision-making.

These insights enable businesses to target customer groups effectively and optimize sales strategies, driving better outcomes and performance.

GitHub Link: https://github.com/sinthiagupta/CustomerInsightPredictor

1. Introduction

Understanding customer behavior is essential for businesses to optimize marketing strategies and improve revenue generation. The **Customer Segmentation and Sales Prediction** system focuses on identifying customer clusters based on spending behavior and predicting future spending trends.

Key goals include:

- **Customer Segmentation**: Grouping customers into clusters based on spending, demographics, and income levels.

- **Sales Prediction**: Forecasting customer spending to aid in inventory management and financial planning.

1.1 Motivation

1. **Actionable Insights:** Efficient customer segmentation and sales forecasting from large datasets.
2. **Resource Optimization:** Targets high-value customer groups and improves planning for promotions and inventory.
3. **User Accessibility:** A Tkinter GUI makes machine learning accessible to non-technical users.

1.2 Challenges

1. **Data Quality**: Handled missing values with mean/mode imputation.

2. **Feature Scaling**: Applied **MinMaxScaler** to normalize data.

3. **Cluster Optimization**: Used Elbow Method and Silhouette Scores for K-Means clustering.

4. **GUI Delays**: Managed heavy computations with progress updates.

1.3 Contribution

1. **Unified Pipeline**: Combines segmentation and prediction for analytics.

2. **Data Handling**: Robust preprocessing for missing values and feature encoding.

3. **Usability**: A user-friendly GUI for model execution and results visualization.

2. Related Survey

Research on **customer segmentation** and **sales prediction** using **K-Means** and **Linear Regression** has been widely explored:

1. Customer Segmentation with K-Means:

- **K-Means Clustering** is commonly used for segmenting customers based on demographics and behavior (Jain, 2010). However, challenges include selecting the optimal number of clusters and handling high-dimensional data.

2. Sales Prediction with Linear Regression:

- **Linear Regression** is a popular method for predicting sales (Zhang & Wang, 2009). While simple and interpretable, it may require careful feature engineering and assumptions about linearity in the data.

These approaches, though effective, often require additional techniques to handle data quality issues and to optimize the models for better accuracy.

---

3. Datasets

The dataset combines three major types of information about customers:

- **Demographics:**
    - Includes attributes like:
        - **Gender** (Male, Female).
        - **Age** (in years).
        - **Annual Income** (in thousands of dollars).
- **Spending Patterns:**
    - Features include:
        - **Spending Score** (a score assigned based on customer behavior and purchasing habits).

    - **Historical Purchases** (total number or frequency of past purchases).

- **Sales Figures:**
    - Captures **total spending** by customers, measured in thousands of dollars ($k), used as the target variable for prediction models.

3.1 Data Preprocessing

Several preprocessing techniques were applied to prepare the dataset for customer segmentation and sales prediction tasks:

1. **Handling Missing Values**:
    - **Numerical Columns**: Missing values in columns like income were imputed using the **mean** of the respective column to maintain consistency without distorting the data distribution.
    - **Categorical Columns**: Missing values in columns such as gender were imputed with the **mode**, representing the most frequent category in the data.

2. **Feature Scaling**:
    - **MinMaxScaler** was used to scale numerical features, such as income and spending scores, to a range between 0 and 1. This step ensures that all features contribute equally to the model, preventing dominance of variables with larger value ranges.

3. **Encoding Categorical Variables**:
    - Categorical variables, such as gender, were encoded using **OneHotEncoder**. This technique creates binary columns (e.g., **Gender_x_Male**, **Gender_y_Male**) for each category, enabling the machine learning algorithms to process them efficiently.

4. **Outlier Detection and Removal**:

- Outliers in numerical features, such as income, were identified using statistical methods (e.g., z-scores) and either removed or capped to ensure they do not skew model training.

5. **Feature Engineering**:

   - New features, such as total spending or average spending per customer, were created from existing columns to improve model performance and provide deeper insights into customer behavior.

These preprocessing steps ensure the dataset is clean, normalized, and ready for analysis, improving the performance of both K-Means clustering and Linear Regression models.

---

4. Methodology

1. K-Means Clustering

- **Objective**: Group customers based on demographics and spending behavior.
- **Process**:
  1. Randomly select kkk centroids.
  2. Assign customers to the nearest centroid.
  3. Recalculate centroids and repeat until stable.
- **Evaluation**: Use **Silhouette Score** for cluster separation and **Inertia** for tightness.

2. Linear Regression

- **Objective**: Predict sales using customer data.

- **Model**: y=β0+β1x1+⋯+βnxny = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_ny=β0+β1x1+⋯+βnxn (sales = intercept + coefficients * features).

- **Evaluation**: Measure performance using **Mean Absolute Error (MAE)**, **R² Score** and **Root Mean Absolute Error (RMAE)** for accuracy

---

4.1 Hardware and Software Requirements

- **Hardware:**
  - **Processor:** Intel i5 or equivalent.
  - **RAM:** 8GB or higher.
  - **Storage:** 500GB HDD or SSD.
- **Software:**
  - **Programming Language:** Python 3.x
  - **Libraries:**
    - Pandas for data manipulation.
    - NumPy for numerical operations.
    - Scikit-learn for machine learning models (KMeans, Linear Regression).
    - Matplotlib/Seaborn for data visualization.
    - Tkinter for building the GUI.
  - **IDE:** Jupyter Notebook or Visual Studio Code.

4.2 Performance Metrics

- **K-Means Clustering:**
  - **Silhouette Score:** Measures cluster separation and cohesion, ranging from -1 to 1 (higher is better).
  - **Inertia:** Measures the sum of squared distances from each point to its assigned centroid (lower is better).

- **Linear Regression:**
  - **Mean Absolute Error (MAE):** Measures the average of absolute errors between predicted and actual values (lower is better).
  - **R² Score:** Indicates the proportion of variance explained by the model (higher is better).
  - **Root Mean Absolute Error (RMAE):** The square root of the MAE, giving error in the same unit as the target variable (lower is better).

5. Results and Analysis

5.1 K-Means Clustering Results

- **Optimal Number of Clusters**: The Elbow Method and Silhouette Score indicated that the optimal number of clusters was **8**.

5.2 Linear Regression Results

- **R² Score**: **0.82**, indicating that 82% of the variance in sales is explained by the model.

- **Mean Absolute Error (MAE)**: **1.85k**, indicating an average error of $1,850 in sales predictions.

- **Root Mean Absolute Error (RMAE)**: **1.36k**, showing the error in the same scale as the sales figures.

5.3 Analysis

- The **K-Means Clustering** effectively grouped customers into meaningful segments, supporting targeted marketing.

- The **Linear Regression** model showed strong performance with a high R² score and low error metrics, making it reliable for sales forecasting.

- These insights provide valuable information for businesses to optimize marketing strategies and inventory planning.

---

6. Conclusions and Future Work

6.1 Conclusions

This project successfully used K-Means Clustering and Linear Regression for customer segmentation and sales prediction. The K-Means algorithm identified three customer segments, enabling targeted marketing, while the Linear Regression model achieved an R^2 Score of 0.82 and an MAE of 1.85k, providing reliable sales forecasts. The Tkinter GUI made the system accessible to non-technical users.

6.2 Future Work

Future improvements could include exploring advanced models like Random Forest or Gradient Boosting for better accuracy, experimenting with other clustering algorithms like DBSCAN, and implementing real-time data processing. Deploying the system as a web application would further enhance its accessibility and scalability.
