SEAS 6414 - Python Applications in Data Analytics
Homework 7

Due Date: March 2, 2024 (10:00am EST)

Instructions: To complete the following task using Python, please download an Integrated Development Environment (IDE) of your choice. Ensure that your solution includes both the written code (input) and its corresponding output. Once completed, upload your solution in PDF format or any other format you prefer. **The questions are worth 50 points each**.

Question 1: Segmentation, Assignment Generation, and Churn Prediction Using Merchant Transaction Activity

Dataset and Objective

- **Dataset:** `homework7_file1.csv`
- **Data Description:** This dataset contains records of merchant transactions, each with a unique merchant identifier, time of transaction, and amount in cents.
- **Objective:** Analyze merchant transaction data to understand business growth and health. Preprocess the dataset for future transactions and generate specific features for each merchant.

Task 1: Feature Generation

Generate the following features for each unique merchant:

- `trans_amount_min`: Minimum transaction amount for each merchant.
- `trans_amount_max`: Maximum transaction amount for each merchant.
- `trans_amount_avg`: Average transaction amount for each merchant.
- `trans_amount_volume`: Total transaction amount for each merchant.
- `trans_frequency`: Total count of transactions for each merchant.
- `trans_recency`: Recency of the last transaction (in days from 1/1/2035).

- `avg_time_btwn_trans`: Average time between transactions (in hours).
- `avg_trans_growth_rate`: Average growth rate in transaction amounts.

Task 2: Merchant Segmentation

- **Goal:** Identify different kinds of businesses in the sample and generate assignments for each merchant using only the given data.
- **Approach:**
  - Use a clustering algorithm to cluster the dataset with the newly created features.
  - Apply silhouette and/or elbow method to determine the optimal number of clusters.
  - Interpret clusters to generate assignments for each merchant.
  - Conduct exploratory data analysis for each cluster to identify and infer different types of businesses.

Task 3: Churn Prediction

- **Background:** Customer retention is a key growth pillar. Churn or customer attrition is defined as customers who have had no transactions within a 30-day period, indicating a "rolling" Monthly Recurring Revenue (MRR) of $0.
- **Goal:** Develop a churn prediction model.
  - Generate binary labels for churn and no churn based on the 30-day inactivity criterion.
  - Use the generated features and any algorithm(s) of your choice for the model.
  - Present metrics from your experiments and feature importance.

Deliverables

- **For Merchant Segmentation:**
  - Cluster assignments for each merchant.
  - Exploratory data analysis report for each cluster.
- **For Churn Prediction:**
  - A developed churn prediction model.
  - Experiment metrics and analysis of feature importance.

Question 2: Unsupervised Learning on Prpoperty Dataset with PCA and Clustering Techniques

Given the comprehensive dataset provided from Zillow, which encompasses various features ranging from architectural details to taxation information, students are tasked to apply unsupervised learning methods to uncover underlying patterns in the data. The assignment will involve two main techniques: Principal Component Analysis (PCA) and Clustering.

Objective

- Use PCA to reduce the dimensionality of the dataset, thus simplifying the data without significant loss of information.
- Apply clustering techniques to identify distinct groups within the Zillow dataset that share similar characteristics.

Tasks

Data Preprocessing:

- **Dataset:** `homework7_file2.csv`
- Normalize the features in the dataset to ensure that the PCA and clustering algorithms function correctly.
- Handle missing values and categorical variables appropriately.

Principal Component Analysis (PCA):

- Implement PCA on the preprocessed dataset.
- Determine the number of principal components to retain by examining the explained variance ratio.

Clustering:

- Choose a suitable clustering algorithm, such as K-means or hierarchical clustering.
- Utilize the silhouette score and/or the elbow method to find the optimal number of clusters.
- Cluster the data based on the principal components derived from PCA.

Cluster Analysis and Interpretation:

- Perform an exploratory data analysis (EDA) on each cluster to understand the properties and characteristics of the different groups.
- Identify and describe the types of properties in each cluster, considering features such as 'yearbuilt', 'roomcnt', 'bathroomcnt', 'zipcode' and others relevant to real estate valuation.

Reporting:

- Present a detailed report of the PCA and clustering process.
- Provide an insightful interpretation of each cluster, supported by statistical summaries and visualizations.
- Discuss the potential implications of your findings for stakeholders in the real estate market.

Deliverables

- A Jupyter notebook containing the PCA and clustering code with detailed annotations explaining each step.
- A written report summarizing the methodology, findings, and interpretations of the clusters.
- Visual aids such as scatter plots, bar charts, and heatmaps to illustrate the characteristics of each cluster.

Additional Notes

- The dataset is expected to contain a mix of numerical and categorical data; proper encoding methods should be applied.
- Clusters should be profiled based on key features that could be useful for property market segmentation.

This assignment will test your ability to apply unsupervised machine learning techniques to real-world datasets and derive actionable insights that can inform business strategy.