Write-up

Overview

The project analyzes the prominent characteristic categories associated with churned customers and customers who decide to continue using a bank's credit card. It operates on a data set consisting of 10,000+ customers, including their age, salary, marital status, and more (nearly 12 features), and indicates whether they have stopped using the card. To run it, simply run cargo run –release on Terminal.

Data Analysis Process:

1. **Graph Construction:** The project begins by reading the data and constructing an undirected graph, where nodes represent customers and connections indicate shared characteristics.
2. **Customer Classification:** Nodes are categorized into churn and not churn groups, and high centrality nodes in both groups are identified using functions from the graph_utils module.
3. **Shared Characteristic Identification:** The top shared characteristic categories and their distribution are printed using functions from the customer module.

Modules Overview:

1. **Main.rs**:
    a. The main function reads the CSV file, extracts information into the Customer Struct (defined in Customers.rs), and constructs an undirected graph using the construct_graph function from Graph_utils.rs. Nodes represent customers, connected when sharing characteristics. The undirected graph is chosen for its suitability in depicting shared characteristics without directional implications.
    b. The Customer Struct is then divided into churned and not churned groups. Centrality is calculated for each node in both groups, identifying nodes with high centrality through the calculate_centrality and identify_high_centrality_nodes functions from Graph_utils.rs.
    c. Finally, the print_top_shared_characteristics function from Customers.rs displays the top 4 shared characteristic categories and their characteristic distributions for each group.
2. **Graph_utils.rs**:
    a. This module handles graph construction and analysis. It imports petgraph and defines functions such as construct_graph that iterates through customer nodes to determine neighbors using determine_neighbor based on shared characteristics. Centrality is calculated using the dijkstra function, and identify_high_centrality_nodes filter nodes with centrality scores exceeding the average times a threshold factor.
    b. The threshold factor, currently set at 1.1 in main.rs, can be adjusted for dataset size and concentration. Lowering it may yield more meaningful results if there's minimal output.
3. **Customers.rs**:
    a. Defines the Customer Struct detailing characteristics, including an embedded struct named OneHotEncoding for categorical variables, imported in main.rs. Functions in this module, such as print_top_shared_characteristics, process high-centrality nodes. find_top_shared_characteristics iterates through neighbors using get_shared_characteristics to record and return the top 4 shared characteristics for each node-neighbor pair.

b. Results are printed and formatted with counts representing shared characteristics, organized by characteristic (e.g., Married, Single, etc.), and further categorized (e.g., Marital Status, Card Type).

```
Churn High Centrality Nodes
Prevalent characteristic categories and their compositions:
Number of Products Purchased, (Total Count: 12 - 2%)
  2: 12 (100%)
Card Type, (Total Count: 144 - 24.3%)
  Blue: 144 (100%)
Education Level, (Total Count: 10 - 1.7%)
  Graduate: 10 (100%)
Age, (Total Count: 182 - 30.7%)
  2: 182 (100%)
Month inactive, (Total Count: 48 - 8.1%)
  3: 48 (100%)
Number of Contacts from Bank (past 12 months), (Total Count: 64 - 10.8%)
  3: 64 (100%)
Marital Status, (Total Count: 132 - 22.3%)
  Married: 132 (100%)

Not Churn High Centrality Nodes:
Prevalent characteristic categories and their compositions:
Education Level, (Total Count: 1980 - 1.8%)
  High School: 74 (3.7%)
  Graduate: 1906 (96.3%)
Month inactive, (Total Count: 6888 - 6.3%)
  3: 1100 (16%)
  2: 5116 (74.3%)
  1: 672 (9.8%)
Income Range, (Total Count: 70 - 0.1%)
  $60K - $80K: 70 (100%)
Number of Products Purchased, (Total Count: 162 - 0.1%)
  4: 84 (51.9%)
  6: 78 (48.1%)
Age, (Total Count: 41146 - 37.4%)
  2: 41146 (100%)
Number of Contacts from Bank (past 12 months), (Total Count: 9528 - 8.7%)
  2: 6250 (65.6%)
  3: 3278 (34.4%)
Card Type, (Total Count: 35744 - 32.5%)
  Blue: 35744 (100%)
Marital Status, (Total Count: 14486 - 13.2%)
  Married: 11830 (81.7%)
  Single: 2656 (18.3%)
```

c. Some categories internally segregate characteristics using ranges, with each range accompanied by a label for better interpretation.

Category	**Labels for Each Characteristic**				
	1	2	3	4	5
	Content of each characteristic Label				
Age	20-30	30-40	40-50	>50	
Mon_w_bank (how many months the customer stayed as a customer)	20-30	30-40	40-50	>50	
Transactions_amount (average dollar amount of transactions on card)	500<	500-1000	1000-1500	1500-2000	>2000
Num_transctions (average number of transactions on card)	<10	10-20	20-30	30-40	>40
Avg_card_utilize (Average Card Utilization Ratio (divide your balance by your credit limit)	<0.1	0.1-0.2	0.2-0.3	0.3-0.4	>0.4

d. Customer.rs also includes tests. Within the test module, two simulated customers are generated, and the functions test_shared_characteristics and test_determine_neighbor are employed to assess the functionality of shared_characteristics and determine_neighbor.

Output Analysis:

The output aligns with our intuition. One noteworthy shared characteristic is the quantity of products customers acquire from the bank. Among customers retaining the card, purchasing more products is a common shared characteristic. This observation is logical, as customers who have already bought numerous items from the bank are intuitively more inclined to keep using the credit card from the same institution. Additionally, a greater percentage of shared characteristics among customers with high centrality in the retaining card group is related to having a higher income. This observation is consistent with the intuition that individuals with more financial resources are more likely to persist in using a credit card.

Citation:

- Data source: Kaggle
    - https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction
- Petgraph and Dijkstra reference: petgraph documentation
    - https://docs.rs/petgraph/latest/petgraph/
- Centrality formulas: Wikipedia, Towards Data Science
    - https://en.wikipedia.org/wiki/Centrality
    - https://towardsdatascience.com/notes-on-graph-theory-centrality-measurements-e37d2e49550a
- Concept clarification, debugging, write-up grammar revision and comment help: ChatGPT
    - https://chat.openai.com/

- Credits: Penny Lin (data set selection), TingYi Wu (debugging support)